Statistical Language Modelling of Dialogue Material in the British National Corpus

A thesis submitted to the University of London
for the degree of Doctor of Philosophy

by

Gordon James Allan Hunter

Department of Phonetics and Linguistics,
University College London.

2004

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Statistical language modelling may not only be used to uncover the patterns which underlie the composition of utterances and texts, but also to build practical language processing technology. Contemporary language applications in automatic speech recognition, sentence interpretation and even machine translation exploit statistical models of language. Spoken dialogue systems, where a human user interacts with a machine via a speech interface in order to get information, make bookings, complaints, etc., are example of such systems which are now technologically feasible.

The majority of statistical language modelling studies to date have concentrated on written text material (or read versions thereof). However, it is well-known that dialogue is significantly different from written text in its lexical content and sentence structure. Furthermore, there are expected to be significant logical, thematic and lexical connections between successive turns within a dialogue, but "turns" are not generally meaningful in written text. There is therefore a need for statistical language modeling studies to be performed on dialogue, particularly with a longer-term aim to using such models in human-machine dialogue interfaces.

In this thesis, I describe the studies I have carried out on statistically modelling the dialogue material within the British National Corpus (BNC) – a very large corpus of modern British English compiled during the 1990s.

This thesis presents a general introductory survey of the field of automatic speech recognition. This is followed by a general introduction to some standard techniques of statistical language modelling which will be employed later in the thesis. The structure of dialogue is discussed using some perspectives from linguistic theory, and reviews some previous approaches (not necessarily statistical) to modelling dialogue. Then a qualitative description is given of the BNC and the dialogue data within it, together with some descriptive statistics relating to it and results from constructing simple trigram language models for both dialogue and text data.

The main part of the thesis describes experiments on the application of statistical language models based on word caches, word "trigger" pairs, and turn clustering to the dialogue data. Several different approaches are used for each type of model. An analysis of the strengths and weaknesses of these techniques is then presented.

The results of the experiments lead to a better understanding of how statistical language modelling might be applied to dialogue for the benefit of future language technologies.

Acknowledgements

I would like to thank the many staff and students of the Department of Phonetics and Linguistics for making me feel so welcome in the Department over the course of my time working there. Particular thanks should go to my friends and office-mates Abbas Haydari, Hyunsong Chung, Bronwen Evans, Catherine Siciliano & Matt Smith, and to Alex Fang for helpful advice on the use of the British National Corpus.

My supervisor, Dr. Mark Huckvale, has been a great source of inspiration and helpful advice over the period of this work, for which I am extremely grateful.

My work has been financially supported by the EPSRC of the UK, through the award of a postgraduate research studentship.

Many friends, too numerous to mention individually, have been a source of great encouragement to me over the period of this work.

Finally, I would like to thank my parents for their patience over the years I have been studying.

Table of Contents

Title Page 1

Abstract 2

Acknowledgements 3

Table of Contents 4

List of Figures 7

List of Tables 8

Main Body of Thesis 12

Chapter 1 Motivation for this study **12**
1.1 Automatic Speech Recognition - Background and Current Status 12
1.2 Language Modelling 21
1.3 Measures of Performance 22
1.4 Aims of this Study 23
1.5 Types of Dialogue and Sources of Data 27

Chapter 2 Language Modelling and Related Work **30**
2.1 Statistical Language Models and N-grams 30
2.2 Adaptive (Dynamic) Models 33
2.2.1 Trigger Pair Models 34
2.2.2 Cache Models 38
2.3.3 Cluster-Based Methods 39
2.2.3.1 Linguistically (Lexically)-Motivated Clustering Methods 40
2.2.3.2 Clustering Methods based on Perplexity 44
2.3 Combining Information from Several Sources 46
2.3.1 Introduction 46
2.3.2 Linear Interpolation 47
2.4 The Maximum Entropy Method 49
2.4.1 Introduction 49
2.4.2 Mathematical Framework : Feature-Based Models and Maximum Entropy 51

Chapter 3 Dialogue & Discourse **57**
3.1 Motivation : What's special about dialogue ? 57
3.2 Dialogue & Discourse : some perspectives from linguistic theory and psycholinguistics 60
3.3 Previous work on modelling dialogue 69
3.4 Aims and Hypotheses for the Remainder of this Study 76

Chapter 4 Dialogue Data in the British National Corpus (BNC) 77
4.1 The British National Corpus (BNC) 77
4.2 Spoken Material (including Dialogue) within the BNC 78
4.3 Descriptive Statistics 79
4.4 Lexical Distribution 81
4.5 Dialogue Reduced-Turns (DRT) Dataset 82
4.6 Simple Statistical Language Models 83
4.6.1 Trigram Models for Ordinary Dialogue Data 83
4.6.2 Trigram Models for DRT Data 86

Chapter 5 Experiments Using Cache-Based Language Models **88**
5.1 Overview 88
5.2 Cache Experiments on Ordinary Dialogue Data from the BNC 89
5.2.1 Comparison of Fixed-Size, Turn-Based and Sentence-Based Caches 89
5.2.2 Variation of the Cache Size 92
5.3 Cache Experiments on DRT Data 95
5.4 Qualitative Observations on Results from Cache-Based Models 97
5.5 Summary 103

Chapter 6 Experiments Using Language Models Based on Trigger Pairs **106**
6.1 Motivation and Overview 106
6.2 Trigger Model Experiments on Ordinary Dialogue Data from the BNC 109
6.2.1 Experiments where the number of triggers per target word was restricted 109
6.2.2 Controlled Experiments Using a Fixed Number of Triggers 115
6.3 Trigger Model Experiments on DRT Data 119
6.4 Comparison of "Best" Trigger Pairs for Dialogue, Text & DRT Data 122
6.5 What Kind of Turn Pairs Benefit Most from the Use of a Trigger Model 131
6.6 Summary 135

Chapter 7 : Experiments Using Language Models based on Clusters **138**
7.1 Overview 138
7.2 Clustering Experiments on Ordinary Dialogue Data – Using Whole Dialogues and a "Mixture of Clusters" Model 139
7.3 Clustering Experiments on DRT Dialogue Data Using a Lexically-Motivated Similarity Metric 141
7.3.1 Dependency of Model Perplexity on Number of Clusters Used 144
7.3.2 Dependency of Model Perplexity on Size of Lexicon Used 149
7.4 Experiments Using an Entropy-Based Clustering Metric 152
7.5 Comparison of Entropy-Based and Lexically-Based Clustering Methods 155
7.6 Qualitative Discussion on the Content of Individual Clusters 157
7.7 What types of turn benefit most from cluster-based modelling ? 159
7.8 Summary 171

Chapter 8 : Discussion, Conclusions and Suggestions for Further Work **173**
8.1 Discussion and Conclusions 173
8.2 Suggestions for Further Work 179

References **182**

Appendices

Appendix A : Some Further Examples of Turn Pairs Showing Large Increases in Probability Through Use of a Cache 198

Appendix B : "Learning on the Job : The Application of Machine Learning within the Speech Recognition Decoder" 205
(Paper presented at 2001 Workshop on Innovation in Speech Processing - WISP 2001 and published in Proceedings of the Institute of Acoustics, Vol. 23 (3), pp 71-79.)

List of Figures

Figure		Page number
1.1	A schematic representation of a typical speech recognition system	20
1.2	A schematic representation of an automated dialogue system	26
4.1	Graph contrasting coverage of material by lexica of various sizes for BNC text and dialogue data	82
4.2	Variation of perplexity of trigram language models (with respect to excluded data) with size of training corpus for text data with corresponding values for models trained on dialogue	85
4.3	Comparison of trigram models for DRT data	87
5.1	Relative perplexity improvement of an interpolated trigram-cache model over a trigram model	95
7.1	Production of clusters and "shadow clusters" for first and second turns of pairs in a "T"-type experiment.	142
7.2	Production of clusters and "shadow clusters" for first and second turns of pairs in an "R"-type experiment.	143

List of Tables

Table		Page number
1.1	Standard speed recognition tasks, listed in increasing order of their difficulty	18
3.1	Some corpora of transcribed dialogues in English	57
4.1	Some summary descriptive statistics for the BNC dialogue material	80
5.1	Summary of results from experiments using cache models.	91
5.2	Comparision of perplexities and interpolation parameters of interpolated trigram-cache models trained and tested on dialogue material from the BNC	93
5.3	Comparison of perplexities and interpolation parameters of interpolated trigram-cache models trained and tested on "TEQ" material from the BNC	94
5.4	Summary comparison of perplexity scores for trigram only (baseline) and interpolated trigram-cache models for DRT data.	97
6.1	Summary of results from experiments using trigger models	112
6.2	Summary results for interpolated trigram-trigger models using a fixed-size window of 500 words.	113
6.3	Summary results for interpolated trigram-trigger models using the current and previous dialogue turn as the window	114
6.4	Summary results for interpolated trigram-trigger models using the current and previous sentence as the window	115
6.5	Variation of perplexities of interpolated trigram-trigger models for dialogue and TEQ data with the number of trigger pairs incorporated into the model. Current and previous sentence used as the window	116

List of Tables (continued)

Table		Page number
6.6	Improvement of perplexities and variation of interpolation of interpolation weights of interpolated trigram-trigger models for dialogue and TEQ data with the number of trigger pairs incorporated into the model. Current and previous sentence used as the window.	116
6.7	Variation of perplexities and interpolation weights of interpolated trigram-trigger models for dialogue data with the number of trigger pairs used in the model. Current and previous turn used as the window.	117
6.8	Variation of perplexities of interpolated trigram-trigger models for dialogue and TEQ data with the number of trigger pairs incorporated into the model. Fixed window of previous 500 words	118
6.9	Improvement of perplexities and variation of interpolation weights of interpolated trigram-trigger models for dialogue and TEQ data with the number of trigger pairs incorporated into the model. Fixed window of previous 500 words.	118
6.10	Summary comparison of perplexity scores for trigram only and interpolated trigram-trigger models for DRT data	121
6.11a	The 28 "best" trigger pairs for text data in the BNC obtained using the top 50000 words in the text vocabulary and a fixed window of 500 words.	124
6.11b	The 28 "best" trigger pairs for text data in the BNC obtained using a restricted text vocabulary of 10000 words and a fixed window of 500 words.	125
6.12a	The 28 "best" trigger pairs for ordinary dialogue data in the BNC obtained using the full 50000 word dialogue vocabulary and a fixed window of 500 words.	127
6.12b	The 28 "best" trigger pairs for ordinary dialogue data in the BNC obtained using a restricted dialogue vocabulary of 10000 words and a fixed window of 500 words.	128
6.13	The 28 "best" trigger pairs for DRT data in the BNC obtained using the full 50000 word in dialogue vocabulary, strictly using the previous turn as window.	130

List of Tables (continued)

Table		Page number
7.1	Comparison of weighted average perplexities of cluster-based, simple trigram and interpolated cluster-trigram language models for DRT data using different numbers of clusters. "F"-type experiment.	146
7.2	Comparison of weighted average perplexities of cluster-based, simple trigram and interpolated cluster-trigram language models for DRT data using different numbers of clusters. "T"-type experiment.	147
7.3	Comparison of weighted average perplexities of cluster-based, simple trigram and interpolated cluster-trigram language models for DRT data using different numbers of clusters. "R"-type experiment.	148
7.4	Comparison of weighted average perplexities of cluster-based, simple trigram and interpolated cluster-trigram language models for DRT data using different numbers of clusters. "O"-type experiment.	149
7.5	Comparison of weighted average perplexities of cluster-based, simple trigram and interpolated cluster-trigram language models for DRT data using different sizes of lexicon. "F"-type experiment.	150
7.6	Comparison of weighted average perplexities of cluster-based, simple trigram and interpolated cluster-trigram language models for DRT data using different sizes of lexicon. "O"-type experiment.	151
7.7	Weighted average perplexities of cluster-based, simple trigram and interpolated cluster-trigram language models for DRT data, where clustering was done using the entropy-based metric. "O"-type experiment.	153
7.8	Summary of cluster sizes and perplexities of cluster-based models, "O" type experiment using the entropy-based metric.	154

List of Tables (continued)

Table		Page number
7.9	Summary of range of perplexities found across clusters : simple trigram models and interpolated trigram-cluster models. "O"-type experiments using the entropy-based clustereing metric.	154
7.10	Average relative reduction in perplexity (with respect to baseline of simple trigram model) of interpolated trigram-cluster model using different numbers of clusters and the two different clustering metrics. "O"-type experiments.	155
7.11	CPU time required per cross-validation rotation on a 2.0 GHz Pentium 4 PC with 512 Mbytes RAM for cluster model experiments with the entropy-based clustering metric.	156

Chapter 1 Motivation for this Study

1.1 Automatic Speech Recognition - Background and Current Status

Speech recognition technology has featured prominently in science fiction stories for many years - from the "speakwrite" machines in George Orwell's *Nineteen Eighty-Four* (1948) to the conversing computers in *Star Trek* and "HAL" in *2001 - A Space Odyssey.*

The principle aim of automatic speech recognition is to develop methods and technology through which instructions and information (data) can be given to machines (including computers), through the medium of speech, in a way which is as robust to adverse conditions and as simple, flexible and natural to human users as is possible. The closer we get to achieving this aim, the simpler human-computer (or, more generally, human-machine) interaction should become (Allen et al, 2001). Users should be able to communicate with devices in a way which will require little specialised training or adaptation of behaviour (see, e.g. Lea, 1980). Automatic speech recognition could provide assistance for deaf or blind people, and for the general public in complete darkness (perhaps removing problems such as finding the right key and getting it in the lock in the dark). They could also facilitate interaction with computers for the very young and the illiterate. Already, speech recognition systems are being used in "speech-to-text" dictation machines and for voice activated dialling in mobile telephones, reducing the need for users to remember and dial so many numbers, and facilitating hands-free operation. Prototypes of systems such as speech to speech language translation machines are also now being made. In many other potential applications, freeing-up the user's hands and eyes to perform other tasks could be greatly beneficial, particularly in situations such as driving a car, flying an aeroplane or performing medical operations. In such situations, the "core task" makes large demands on the user's hand-eye co-ordination and it is desirable that auxiliary tasks cause the minimum distraction from the core.

However, in order to achieve this goal, the speech recognition system will have to become very robust and flexible - able to cope with both a very large vocabulary and

a wide range of speakers, adapt to the different emotional states or health conditions of the speaker, noisy environments, hesitations, false starts and non-words on the part of the speaker. Even the best modern systems fall short of these targets. For example, a system which achieves a high recognition performance once trained to the normal speaking voice of a particular user may show a greatly inferior performance if that same user develops a cold. It is widely considered that the practical application of automatic speech recognition will only become very widespread once their performance becomes comparable with that of humans under ordinary listening environments (Lippmann, 1997)

In fact, scientists and engineers have been attempting to develop speech recognisers since at least 1930, when the Hungarian Tihamér Nemes unsuccessfully applied for a patent for an automatic speech transcription system making use of the soundtrack on a ciné film (Kohonen, 1988). Speech recognition has been a topic of serious research since the 1950's, when simple recognisers were constructed which showed some success over restricted domains, such as recognition of spoken words for single digits (Waibel & Lee, 1990). Indeed, experiments in this field were first carried out at UCL during the mid 1950's (Fry & Denes, 1955, Fry 1959, Denes, 1959). However, only limited progress was made, despite substantial investment of time, resources and money during the 1960's, leading to many becoming highly cynical regarding the future of automatic speech recognition (see, e.g. Peirce, 1969).

Computer technology advanced rapidly (and became much cheaper) during the 1970's and '80's, and together with improved knowledge of speech science, innovations in speech processing algorithms and language modelling, significant advances were made in speech recognition over that period. The first large vocabulary continuous speech recognition system to achieve a reasonable recognition performance was the *Harpy* system, developed in the mid to late 1970's (Klatt 1977 , Lowerre & Reddy 1980).

Currently, as noted above, speech recognition technology is finding practical commercial applications in areas which are familiar to the general public. However, this does not mean that all problems of speech recognition have already been solved. These modern applications tend to be restricted to the most straightforward

conditions, where high recognition performance is simplest - limited vocabularies, often conditioned to just a single speaker, for speech read slowly and clearly from a text, or dictated carefully, in a quiet, noise-free environment. Even state-of-the-art research systems tend to show much poorer performance if required to deal with many different speakers, spontaneous speech with near unlimited vocabulary, hesitations and false starts, perhaps a degraded-quality speech signal (e.g. received over a telephone - often called "telephone speech") and possibly in noisy conditions.

Although many of the problems they highlighted have subsequently to a large extent been overcome, Waibel & Lee (1990) noted that the following aspects of a speech recognition problem would affect the likely success of a recognition system applied to it :

- whether the speech consists of isolated word or is connected or continuous.
- how large the potential vocabulary is.
- task & linguistic constraints on the input speech - will it necessarily consist of syntactically well-formed, semantically meaningful, statistically plausible sentences ?
- whether the system is designed to be used with a single speaker, can adapt to new speakers, or is to perform independent of the speaker.
- potential for acoustic confusability or ambiguity amongst words in the vocabulary.
- variability of conditions, including noise.

A more modern list of factors should include considerations of the quality of articulation, the consistency and prosody of the speech, the coherence of the topic of the speech and how well the speech input matches the assumptions made by the system (for example, strong regional accents and dialects, or input of non-British varieties of English will tends to cause problems for a system expecting speech in standard British English with a "neutral" accent).

Continuous speech recognition (CSR) is normally much more difficult than the recognition of isolated words - to the extent that the main problems of recognising

isolated words are now generally considered to have been solved. Two key problems which occur in the former but not the latter are the determination of word boundaries (for example, a speech signal for "youth in Asia" may be very much like one for "euthenasia", or one for "antelope" like one for "ant elope" - see Calvin & Hobbes cartoon below) and the fact that words in spontaneous connected speech often have poorer articulation and stronger co-articulation than is the case for the same words spoken separately (e.g. "did you" becomes "didya").

Typically, the performance of a speech recognition system - in terms of both accuracy and efficiency - deteriorates as the size of the permitted vocabulary increases. With very large vocabularies, if the system is based on templates of complete words rather than smaller units such as phonemes, it becomes impracticable to perform exhaustive searches for candidate words, or even to acquire enough data to train the system adequately.

Furthermore, in continuous speech, many word sequences which a system might otherwise propose can be ruled out on the basis of syntactic or semantic implausibility (e.g. The famous pair of sentences proposed by Chomsky (1957) "Furiously sleep ideas green colourless" is syntactically bad, whilst "Colourless green ideas sleep furiously" is syntactically well-formed, but semantically implausible), or on the basis of statistical improbability ("This is quite a typical sentence" is more likely to occur than "Utterances comprised of a multiplicity of individually improbable lexemes are consequentially proportionately uncommon"). Pereira (2000) has recently evaluated the relative probabilities of Chomsky's sentences using a statistical language model, finding that the well-formed sentence – despite being implausible – is many times more likely than the ill-formed example.

Even a single speaker can say a given word in many different ways, according to the word's role in a sentence, desired emphasis, the speaker's emotional state and condition of health (as noted previously, the same speaker can sound quite different - both to other people and to a automatic speech recognition system - with a cold). The problem becomes compounded when a system has to be able to cope with several (or even many) different speakers. A distinction is made between a *speaker dependent* system and a *speaker independent* one. In the former case, the system undergoes some additional training to become accustomed to each new speaker it is exposed to. In contrast, a speaker independent system will only be trained once - with speech data from a variety of speakers. However, such training cannot possibly represent all the variability which occurs amongst all the speakers to which the system may be exposed during use, so a speaker independent system tends to have an inferior performance to an otherwise similar speaker dependent one. However, most modern systems tend to be designed with the ability to adapt to a new speaker not encountered in the original training.

Even systems designed for the recognition of a limited vocabulary may be prone to errors if there is the potential for acoustic confusability or ambiguity between the words which the system is designed to recognise. This is unlikely to be a serious problem if the words are all quite distinct -for example, a recognition system designed to speed-up the process of a user inputting a utility-meter reading over the telephone primarily has to recognise the digits "zero" (and its alternative names) to "nine", which all sound quite different. However, this is not always the case. For example (Waibel & Lee, 1990), the names for the letters "B", "P", "D", "G", "C" , "Z" (in American English), "V", "T" and "E" are all quite similar phonetically (several of these only differing from one another by a single phonetic feature such as voicing, place or manner of articulation).

At the risk of stating what may appear obvious, automatic speech recognition systems tend not to perform as well in noisy environments as they do in quiet situations. In fact, any factor which degrades the quality of the input speech signal is likely to have an adverse effect on the system's performance. More than one person speaking at a time, the reduction in bandwidth due to transmission over a telephone system or use

of a poor quality microphone may all reduce the performance of a recogniser. (Some authors - e.g. Lowerre & Reddy (1980) - have suggested that the degradation of the signal due to telephone transmission can lead to an increase in the error rate by a factor of 3 to 4.) Both a poor signal-to-noise ratio and mismatches between the data used for training the system and that on which it is used are likely to reduce the system's performance. In fact, if a system is trained on speech obtained using poor quality equipment, the use of higher quality microphones, etc. during its application may even reduce the success of the system's attempts at recognition ! Variability, rather than just lack of quality, can cause problems.

Summary details of various "standard" speech recognition tasks, listed in increasing order of difficulty, are shown in table 1.1 overleaf. The problems associated with many of these have by now essentially been overcome, and vocabularies of 65000 words or more are now quite common. The "speech understanding" tasks require that the essential *meaning* of an utterance be identified correctly - perhaps for an application such as controlling a mechanical device - whereas the "speech recognition" tasks require the correct identification of the actual words spoken without necessarily interpreting what they mean. The two types of task are therefore somewhat different.

Task	Style of speech	Vocabulary size (words)	Restrictions on usage	Language	Speaker	Environment
Recognition of isolated words	Isolated words	10-300	Very limited	Very restricted vocabulary	Cooperative	Any
Restricted connected speech recognition	Connected speech	30-500	Rather limited	Restricted language of "commands"	Cooperative	Quiet room required
Restricted speech understanding	Connected speech	100-2000	Full usage	English-like	Not un-cooperative	Any
Restricted dictation machine	Connected speech ?	1000-10000	Limited	English-like	Cooperative	Quiet room
Unrestricted speech understanding*	Connected speech	Unlimited	Unlimited	Normal English	Not un-cooperative	Any
Unrestricted connected speech recognition	Connected speech	Unlimited	Unlimited	Normal English	Not un-cooperative	Quiet room ?

Table 1.1

Standard speech recognition tasks, listed in increasing order of their difficulty.
Adapted from Reddy (1976).
(A "not un-cooperative" speaker will not try to confuse the system, but will not try too hard to help it either. In contrast, a cooperative speaker will speak slowly & clearly and will be willing to repeat or spell out words causing problems.
*The "unrestricted speech understanding" task allows use to be made of task-specific information, unlike the unrestricted speech recognition problem).

Although it is still true that Large Vocabulary Continuous Speech Recognition (corresponding to the "Unrestricted connected speech recognition" task in the table above) is the most difficult problem, much progress has been made in this field over recent years. Average word error rates below 10 % for speaker independent systems are now obtainable, and below 5% for speaker-dependent systems, even if the system is just exposed to around 3 minutes of speech from each new speaker (Young, 1996).

Most modern Large Vocabulary Speech Recognition systems work using statistical pattern recognition techniques. These techniques were first applied to speech recognition by Baker (1975) and Jelinek (1976), and developed by Bahl, Jelinek & Mercer (1983). These articles revolutionised automatic speech recognition and the principles they pioneered - to be discussed in Chapter 2 - are still at the core of recognition systems today.

A schematic diagram of a typical speech recognition system is shown in Figure 1.1 overleaf. The speech is input to the system via the microphone on the far left. (In practice, the speech can be pre-recorded.) The pre-processing system includes analog to digital (A to D) conversion, sampling and Fourier and "cepstral" analysis of the speech signal. A more complete discussion of the front-end processing is given in Young (1996). The language model, based on statistical analysis of transcriptions of spoken English, is used to predict the *a priori* probability of any specified word sequence. The acoustic model is based on a pronunciation dictionary of English, and is used to predict what the speech signal for that word sequence should be like. The role of the decoder is to propose possible word sequences which would be compatible with the input speech signal and, based on information from the language model and acoustic model, decide how probable it is that this proposed word sequence is the correct interpretation of the speech signal, and assess the relative merits of the various sequences proposed. In the terminology of probability theory, the language model gives a *prior probability* $P(\underline{\mathbf{w}})$ for a proposed word sequence $\underline{\mathbf{w}}$. The acoustic model then calculates the likelihood or conditional probability for the given acoustic signal $\underline{\mathbf{a}}$ having been generated given the particular word sequence $\underline{\mathbf{w}}$. This is carried out using composite Hidden Markov Models (HMMs) (Bahl, Jelinek & Mercer 1983) based on simple HMMs for phonemes and information in the pronunciation dictionary (Young 1996).

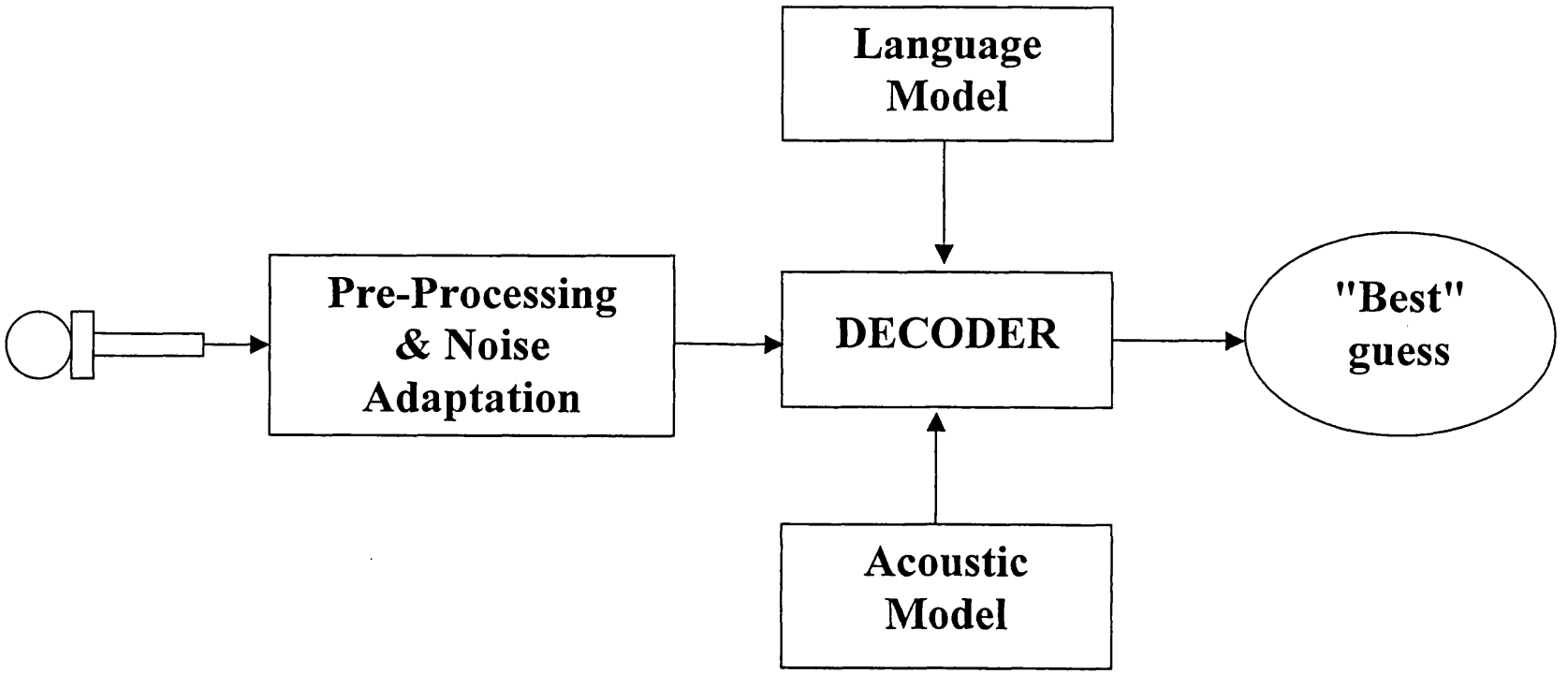


Figure 1.1
A Schematic Representation of a Typical Speech Recognition System.

1.2 Language Modelling

The language model is clearly a key part of a speech recognition system (or "recogniser"). The purpose of a language model is to estimate the overall probability of an utterance and thence, by decomposing the utterance into its composite words, enable the system to predict the next word, w_n, in the utterance on the basis of probabilities, given the sequence of words $W_{1,n-1} = \{w_1, w_2, w_3, \ldots, w_{n-1}\}$ preceding it. The majority of systems do this probabilistically : estimating the probability that the next word is w_n given the preceding sequence $W_{1,n-1}$, i.e they estimate $P(w_n \mid W_{1,n-1})$ (Note that this notation allows more flexibility than the simpler $\underline{w}$, which represents the complete sequence $\{w_1, w_2, w_3, \ldots, w_{n-1}\} \equiv W_{1,n-1}$. We can use $W_{r,s} = \{ w_r, w_{r+1}, w_{r+2}, \ldots, w_{s-1}, w_s \}$ to denote any part of the sequence.)

A commonly-used, simple but effective method for estimating this probability is the use of an N-gram model (Young, 1996). In such a model, it is assumed that the new word only depends on the (N-1) words which immediately precede it, so that $P(w_n \mid W_{1,n-1}) = P(w_n \mid W_{n-N,n-1})$, where $W_{n-N,n-1} = \{ w_{n-N}, w_{n-N+1}, \ldots, w_{n-1}\}$. In practice, since the complete set of possible N word sequences (even restricting the permitted vocabulary to common words) grows very rapidly with N, most language models used in practice are based on trigrams (N=3) and/or bigrams (N=2). Although the principle of N-gram models appears rather crude and they do suffer from several drawbacks - such as failure to take any account of longer-range syntactic or semantic dependencies between words (Bod, 2000, Brill et al, 1998)- they have probably been the most successful single type of language model used in the history of automatic speech recognition. This is particularly true for languages with a strict word order, such as English, since an N-gram will encode syntactic, semantic and pragmatic information relating to a word's nearest neighbours - in the absence of more detailed information, the words most strongly-connected with the word currently of interest - without requiring detailed linguistic rules of the language, such as a grammar, explicitly incorporated into the model (Young, 1996). Indeed, despite their drawbacks, an N-gram model is at the core of most practical speech recognition systems. Despite many attempts to better-use linguistic information in language models, these attempts have made limited improvements over N-gram models (Brill et al, 1998). A more detailed discussion of N-gram models and their strengths and weaknesses, and of other statistical language models, will be made in Chapter 2.

Thus, in summary it can be said that most modern successful automatic speech recognition systems use statistical language models, usually based on word N-gram statistics. This emphasises a "speech and language engineering" approach, where good performance is the most important issue, rather than taking an approach which makes good use of linguistic knowledge or tries to model human speech recognition realistically (Huckvale, 1998).

1.3 Measures of Performance

One measure of the difficulty of a speech recognition task on a sequence of n words $\underline{w} = (w_1, w_2, w_3, \ldots, w_n)$ is the LogProb (or Entropy): $LP(\underline{w}) = (-1/n) \log_2 (P(\underline{w}))$ (Kuhn & De Mori 1990).

However, Bahl et al (1983) have suggested that the perplexity, $S(\underline{w})$ or $PP(\underline{w})$:

$$S(\underline{w}) = 2^{H(\underline{w})} = 2^{LP(\underline{w})} = PP(\underline{w})$$

is a better measure of difficulty, which correlates well with error rate. Using Information Theory, Shannon (1951) showed that the maximum entropy possible for a task for which there are N possible sentences of average length L is $(1/L) \log_2(N)$. Hence, in this situation, the maximum possible perplexity is $N^{(1/L)}$ and so the perplexity can be regarded as the average number of words possible at any given point, i.e. the difficulty of the speech recognition task is equivalent to one where the language has S equally probable words. Alternatively, the perplexity S can be interpreted as the reciprocal of the average probability (as calculated by a language model) per word for a given correct word sequence. The better the language model, the stronger its predictive power, and hence the higher the probability assigned to a "good" or "correct" word sequence. Thus, good language models would be expected to give lower perplexities than bad language models (Kuhn & De Mori 1990).

An alternative way of evaluating the performance of a recognition system is to find its Word Error Rate (WER) - the proportion of words it identifies incorrectly - when performing some standard recognition task on some standard data.

1.4 Aims of this Study

Although attempts to improve language models by making use of more detailed linguistic information have to date met with limited success compared with the cruder N-gram models (Brill et al 1998), intuitively it would seem that a successful, comprehensive model of natural human language *should* contain a lot of linguistic knowledge - both syntactic and semantic - and knowledge about the world. A native speaker of English will normally find it straightforward to distinguish between a sentence which is (a) syntactically badly-formed (e.g. "Furiously sleep ideas green colourless"), (b) one which is syntactically valid but semantically meaningless (e.g. "Colourless green ideas sleep furiously"), (c) one which is O.K. both syntactically and semantically, but is unusual (e.g. "Utterances comprised of a multiplicity of individually improbable lexemes are consequentially proportionately uncommon") and (d) a "normal" sentence (e.g. "This is quite a typical sentence"). However, an automatic speech recognition system might have trouble distinguishing between these - particularly between cases (a) and (b) if the individual words are not too obscure, and between (a), (b) and (c) if the words are less common. As an illustration, the following are examples of an actual recognition system ("*ABBOT*", Hochberg et al, 1995, combined with a decoder developed at UCL) listing what it believes are its "best guesses" (or "hypotheses") at recognising an utterance which is actually "We are going to Paris.". The first column is a marker indicating whether the "guess" is correct as far as it goes (based on knowledge of the correct answer), the second is the name of the file where the "word lattice" for that hypothesis is stored, the third field is a "node number" which can broadly be interpreted as a discrete time counter (how far through the utterance the recogniser has progressed). The fourth field represents an overall log-probability score for the hypothesis (the less negative, the more likely). The numbers in brackets after each word of the hypothesis are normalised log-probability scores for that word.

At a relatively early point (time step 15) in the utterance :

0 sent010.lat.1 15 -3.139 WE(1.500) GO(1.962) TO(-0.466)
0 sent010.lat.1 15 -3.171 WE(1.500) GOING(1.939) TO(-0.466)
0 sent010.lat.1 15 -3.219 WE(1.500) GO(1.962) THE(0.370)
0 sent010.lat.1 15 -3.865 WE(1.500) BOTH(0.858) THE(0.342)
0 sent010.lat.1 15 -4.163 WE(1.500) GOING(1.939) THE(0.370)
1 sent010.lat.1 15 -4.291 WE(2.110) ARE(-1.671) GOING(1.939) TO(-0.466)
0 sent010.lat.1 15 -4.432 WE'LL(0.875) GOING(1.939) TO(-0.466)
0 sent010.lat.1 15 -4.460 WE(1.830) DON'T(-2.173)
0 sent010.lat.1 15 -4.508 WE(1.830) DON'T(0.236) THE(0.342)
0 sent010.lat.1 15 -4.586 WE(1.830) DON'T(-0.569) A(0.208)
0 sent010.lat.1 15 -4.700 WE'VE(0.991) GOING(1.966) TO(-0.466)
0 sent010.lat.1 15 -4.753 WE'RE(0.991) GOING(1.966) TO(-0.466)
0 sent010.lat.1 15 -4.842 WE(1.500) GO(1.962) TO(-0.674) A(0.208)
0 sent010.lat.1 15 -5.000 WE(2.110) HAVE(-1.697) GOING(1.966) TO(-0.466)
0 sent010.lat.1 15 -5.063 WE'D(1.698) GOING(1.966) TO(-0.466)
0 sent010.lat.1 15 -5.074 WE(2.279) A(-0.060) BOTH(0.858) THE(0.342)
0 sent010.lat.1 15 -5.083 WE(1.500) GOING(1.939) TO(-0.674) A(0.208)
0 sent010.lat.1 15 -5.126 WE'LL(0.875) BOTH(0.858) THE(0.342)
0 sent010.lat.1 15 -5.201 WE(2.110) ARE(-1.671) BOTH(0.858) THE(0.342)
0 sent010.lat.1 15 -5.250 WE(2.279) A(-0.060) GOING(1.939) TO(-0.466)

and later on (time step 18) ...

0 sent010.lat.1 18 -5.804 WE(1.500) GO(1.962) THE(0.370) PRESS(0.565)
0 sent010.lat.1 18 -6.201 WE(1.500) GO(1.962) THE(0.370) PARIS(2.987)
0 sent010.lat.1 18 -6.213 WE(1.500) GO(1.962) TO(-0.466) PARIS(2.987)
0 sent010.lat.1 18 -6.246 WE(1.500) GOING(1.939) TO(-0.466) PARIS(2.987)
0 sent010.lat.1 18 -6.321 WE(1.500) GO(1.962) TO(-0.466) PRESS(0.565)
0 sent010.lat.1 18 -6.419 WE(1.500) GOING(1.939) TO(-0.466) PRESS(0.565)
0 sent010.lat.1 18 -6.468 WE(1.500) BOTH(0.858) THE(0.342) PRESS(0.565)
0 sent010.lat.1 18 -6.616 WE(1.500) GO(1.962) TO(-0.466) PRINT(-0.432)
0 sent010.lat.1 18 -6.696 WE(1.500) GOING(1.939) THE(0.370) PRESS(0.565)
0 sent010.lat.1 18 -6.714 WE(1.500) GOING(1.939) TO(-0.466) PRINT(-0.432)
0 sent010.lat.1 18 -6.846 WE(1.500) BOTH(0.858) THE(0.342) PARIS(2.987)
0 sent010.lat.1 18 -7.094 WE(1.830) DON'T(0.236) THE(0.342) PRESS(0.565)
0 sent010.lat.1 18 -7.145 WE(1.500) GOING(1.939) THE(0.370) PARIS(2.987)
0 sent010.lat.1 18 -7.301 WE(1.500) GO(1.962) THE(0.370) PRINT(-0.432)
1 sent010.lat.1 18 -7.365 WE(2.110) ARE(-1.671) GOING(1.939) TO(-0.466)
PARIS(2.987)
0 sent010.lat.1 18 -7.484 WE(1.830) DON'T(-2.173) PARIS(2.987)
0 sent010.lat.1 18 -7.490 WE(1.830) DON'T(0.236) THE(0.342) PARIS(2.987)
0 sent010.lat.1 18 -7.507 WE'LL(0.875) GOING(1.939) TO(-0.466) PARIS(2.987)
0 sent010.lat.1 18 -7.538 WE(2.110) ARE(-1.671) GOING(1.939) TO(-0.466)
PRESS(0.565)
0 sent010.lat.1 18 -7.601 WE(1.830) DON'T(-0.569) A(0.208) PARIS(2.987)

These examples show that a relatively modern recogniser with a trained language model - in this case based on utterances totalling 80 million words from the British

National Corpus (Burnard, 1995) - can propose syntactically ill-formed sentence fragments as interpretations of a syntactically well-formed, meaningful utterance, despite getting many individual words either correct or nearly correct.

It would therefore seem desirable to devise a recognition system which takes the best from both approaches - capitalising on the successes of N-gram based models, whilst trying to make better use of linguistic and contextual information. There have been several attempts at this already (e.g. Brill et al. 1998, Rosenfeld 1996, 2000a), which will be discussed in more detail in Chapter 2. The initial phase of the present study tried to do this by using machine learning principles within the decoder - to investigate how features based on co-occurrences of certain word classes (based on parts of speech) within a hypothesis correlated with whether the hypothesis was or was not "correct so far" (Huckvale & Hunter, 2001, see Appendix B). This showed some improvement in performance over more conventional methods.

There is some evidence which supports the intuitively plausible view that information from earlier sentences should reduce uncertainty over the next word (Zhang, Black & Finch 1999). This could be particularly useful in a practical application to situations such as "dialogue systems" - where a dialogue takes place between a human and a machine (computer) (Allen et al, 2001). An example of a situation where this could be useful is in automated enquiry systems - for instance, an automated telephone "receptionist", such as the "How may I help you ?" system being trialled by AT&T in the USA (Gorin et al, 1997, 2001). This system is an attempt to reduce the frustration caused to customers by the familiar automated "push-button" menu-based call reception systems in common use today ("To make a payment press 1, for bill enquiries press 2, ... , for other options press 0"). Such systems are a cause of much annoyance to customers and frequently result in the customer hanging-up rather than persevering with the call. A dialogue system such as "How may I help you ?" attempts to reduce the irritation caused to customers, whilst still avoiding the expense of having a human operator as the initial point of response, by replacing the "push-button menu" with a speech recognition system. This responds to keywords and phrases in the customer's utterances, making decisions about what question to ask the customer next, or where to re-direct the customer's call. The aim of this system is to provide automated services using a *natural* spoken dialogue system - where the use of *natural*

indicates that the system has to respond to what the human user *actually* says, rather than what the system would like the user to say (Gorin et al, 1997). In this situation, it can no longer assume that the user is cooperative or even "not uncooperative". Further possible applications of dialogue systems include transcription of business or political meetings (e.g. "Hansard"), or of legal court proceedings – where a verbatim record of what has been said by several speakers is required – and real-time (or near-real-time) translation systems (such as VERBMOBIL (Wahlster 1993, 2000)) to facilitate spoken communication between two people who have different first languages.

A schematic representation of a typical dialogue system is shown in figure 1.2 below.

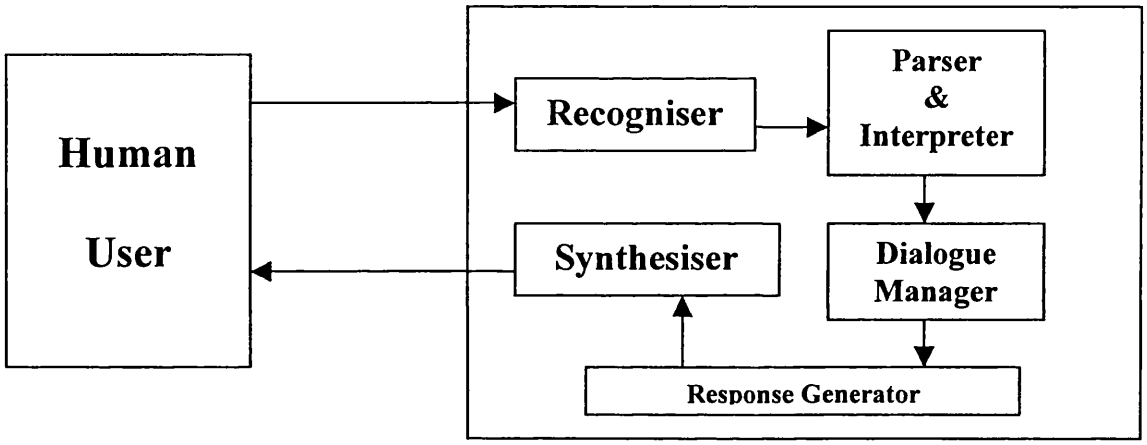


Figure 1.2 A schematic representation of an automated dialogue system. The language model forms part of the unit described as the "recogniser" (After Young, 2000)

The majority of earlier work on spoken dialogue systems have concentrated on either developing a system which performs well if tailored to carrying out simple tasks in a specific, restricted area, or on developing strategies for specific tasks such as requesting clarification (Chu-Carroll, 1999). Many such systems depend heavily on the spotting of keywords or other lexical relations, "semantic parsing" (rule-based mapping from the recogniser output to semantic labels) and "context-dependent semantic decoding" (mapping from these semantic labels to actions) (Potamianos, Riccardi & Narayanan, 1999). These are believed to be appropriate for tasks where

there are few possible actions for the system to take, and most key instructions are likely to be expressed (by the user) using one of a small set of short phrases (or phrase fragments) of low perplexity (Potamianos et al 1999). An example of such a situation is a system for making travel reservations. However, in many other situations, such as an automated telephone receptionist (e.g. "How may I help you ?"), there are many possible actions for the system to take, but the user may not be very co-operative and information given by the user may consist of high perplexity phrases (or phrase fragments). In such cases, it is important to study the mapping from the user's speech input to the machine's identification of the correct task statistically.

Broadly speaking, there have been two quite different approaches to dialogue systems for human-computer interaction - one attempting to develop general computational models for all types of dialogues, the other concentrating on developing systems for specific applications in limited area, using "domain knowledge" related to the precise tasks for that problem. The former approach has the disadvantages of being extremely computationally intensive and possibly "too general" - it may over-emphasise flexibility and so incorporate features not required in the application to which it is being put. The latter approach, on the other hand, may show impressive performance within its limited domain, but may not be easy to generalise or change to deal with different situations (Dahlbäck & Jönsson, 1999).

The focus of the remainder of this study is the investigation of statistical relationships between consecutive dialogue turns and to study how such knowledge can affect the estimated probabilities of proposed word sequences and the performance of language models. The success of the models developed for dialogue will, where possible, be compared with the closest possible corresponding models for written text.

1.5 Types of Dialogue and Sources of Data

The types of dialogue which are most likely to be of interest in Human-Computer Interaction fall into two categories. So-called "task oriented dialogues" (Dahlbäck & Jönsson, 1999, Grosz 1977) relate to situations where both parties (two people or one person and the machine) are trying to cooperate in order for one of them to complete a task. For example, one party may be trying to explain to the other how to do

something, where the "explainer" is initially unsure how much the "explainee" already knows or needs to know. It may be necessary to break the overall task (and associated goal) into sub-tasks and sub-goals. An example of this is the so-called "Map task" (Anderson et al, 1991, Carletta et al 1997), where one party is trying to explain to the other how to get from one place (A) on a map to another (B). The problem can be made more difficult if the two parties do not have identical maps - they will have to negotiate over landmarks which are marked on both maps.

E.g. (Fictious)
(Director)"Turn left at the post office."
(Directee) "I don't know where the post office is."
(Director) "If you continue down the main road for about 200 metres after the church, it should be on the left hand side."
(Directee) "O.K. I know where the church is."

A rather different type of dialogue is a "question answering" dialogue (Dahlbäck & Jönsson, 1999), where the responder cannot be assumed to have the same goal as the questioner, but just has the job of responding to the questioner's questions. An example of this would be a travel enquiry helpline,

e.g. (Fictious)
(Questioner) "What time is the last train from London to Birmingham on Saturday 23rd ?"
(Responder) "From Euston, 23:20, or from Marylebone 22:55"
(Questioner) "What time does each of these arrive in Birmingham ?"
(Responder) "The Euston train reaches Birmingham New Street at 01:15, the Marylebone train reaches Birmingham Midland Road at 01:36."
(Questioner) "How much is the return fare, to come back on Sunday 24th ?"
(Responder) "18.50 going from Marylebone and returning there, 26 pounds from Euston."
(Questioner) "Why the difference ?"
(Responder)"The cheaper fare is only valid on Chiltern Railways, going from Marylebone. The other fare gives you a choice of routes."

There are several corpora of transcribed speech (some also include the actual speech recordings) which could be used as data for such an investigation as this. The British National Corpus (Burnard, 1995) contains a section of transcribed speech, including 672 dialogue files totalling 7760753 words. The Map Task Corpus (HCRC 2001) contains 128 dialogues, all involving one speaker trying to direct the other to a specific destination with reference to a map. The Bramshill corpus (LDC 2001) consists of approximately 600 000 words from a series of 10 minutes conversations, in each of which British Police Officers are describing a series of photographs to each other and discussing them. Clearly, the latter two corpora relate to rather restricted topic domains. There are other large corpora of transcribed dialogue each on a variety of topics - for example, the Switchboard (Godfrey & Holliman 1997), CallHome (Kingsbury et al. 1999) and CallFriend (Canavan & Zipperlen 1997) corpora - but most of these are of American rather than British English which, to be used in the training of a dialogue system for British English, would require different language and acoustic models (different training corpus, pronunciation dictionary, etc.). The British National Corpus has the additional advantage of containing a large body of text material in modern British English, enabling comparisons to be made between textual and dialogue data and between language models for each. Thus, the British National Corpus will be used as the source of data for the remainder of this study.

Chapter 2 Language Modelling and Related Work

2.1 Statistical Language Models and N-grams

Acoustic modelling and processing of speech signals are in themselves not sufficient for achieving satisfactory performance from an automatic speech recognition system applied to large vocabulary problems. Humans rely on many cues and pieces of information which are not purely acoustic when processing speech - syntactic, semantic, pragmatic, dialogue information and (frequently) knowledge about the speaker (Waibel & Lee, 1990, Chapter 8). With the probable exception of "knowledge about the speaker", the process of modelling (or attempting to model) these aspects of the speech recognition process - whether in humans or machines - is called language modelling. Of course, some of these fields, such as pragmatics - which require a higher-level understanding of the language - are harder to model than others. As noted in the previous chapter, the majority of successful automatic speech recognition systems use statistical language models - where the syntactic, semantic and pragmatic information is implicitly encoded in statistics relating to the occurrences of word sequences - based on N-grams.

A statistical language model is essentially a probability distribution, $P(\underline{s})$, over the set S of all possible sentences, utterances or word sequences, $\underline{s} \in S$ (Rosenfeld, 2000b) . Typically, a statistical language model is constructed from data, such as a text corpus - e.g. the British National Corpus (Burnard, 1995) - then used as a "prior" in a system such as a Bayesian classifier to predict the probability of various word sequences given other information, such as an acoustic signal and/or a "dialogue history" in the context of automatic speech recognition.

Given some additional information $\underline{a}$, such as an acoustic signal, we want to find the word sequence, $\underline{s}^*$, which maximises $P(\underline{s} \mid \underline{a})$. By Bayes' theorem,

$$P(\underline{s} \mid \underline{a}) = (P(\underline{a} \mid \underline{s}) . P(\underline{s})) / P(\underline{a})$$

Thus, if $\underline{a}$ is known, $\underline{s}^* = \arg\max (P(\underline{s} \mid \underline{a})) = \arg\max (P(\underline{a} \mid \underline{s}) . P(\underline{s}))$, where arg max denotes locating the value (of $\underline{s}$ in this case) which maximises the expression in brackets. $P(\underline{a} \mid \underline{s})$ is computed by the acoustic model and is the (estimated) probability of $\underline{a}$ being the acoustic signal if $\underline{s}$ is known to be the word sequence.

However, the process of constructing and using the language model in an optimal way is far from trivial. The model would ideally encapsulate sufficient information to enable $\underline{s}$ to be determined from $\underline{a}$ with complete certainty (no errors). Unfortunately, the number of theoretically possible sentences is infinite and any model is going to contain simplifications and approximations. Successful attempts at language modelling are always going to involve some compromise between simplicity (and computational convenience) and better representation of linguistic phenomena. Generally speaking, a more complex model should be better able to account for dependencies between words - particularly words which are further apart in an utterance - but will be less convenient both computationally and from the point of view of obtaining sufficient training data.

N-gram models (Bahl et al 1983, Jelinek 1990) - where it is assumed that the probability of the current word only depends on the immediately preceding (N-1) words - are amongst the simplest and most commonly-used language models. However, for N small (in practice, often N=2 or N=3 are used), the model is very crude and fails to take account of longer-range dependencies between words. Furthermore, such a model may assign relatively high probabilities to word sequences which are nonsensical or ungrammatical, provided that such a sequence does not violate any short-range restriction (Rosenfeld 1996). They also only take account of a word's position in a sequence (sentence or utterance) rather than its linguistic function. In contrast, if a larger value of N is used, the problem of an excessively large number of possible sequences arises. If our model uses a vocabulary of V distinct words, then there are V^N N-gram sequences which, in the absence of additional information, are in principle possible. If V is of the order of several thousand, the number of possible N-grams grows extremely quickly with N. Acquiring and processing sufficient data to obtain reliable estimates of sequence probabilities will be a very serious problem (Bellegarda, 2000). It is generally believed that empirical estimates of probabilities based on 5 or more observations of the occurrence of a particular word sequence

within a training corpus are likely to be reasonably reliable. However, many possible sequences (particularly if N is large and the training corpus is relatively small) will occur fewer than 5 times in the training corpus. Some sequences, despite being plausible English word sequences, may not occur at all (Witten & Bell, 1991) - even for small values of N. Estimates of the probabilities of such sequences calculated directly by counting their occurrences within the training corpus will therefore not be reliable. Special estimation techniques, sometimes called "smoothing", are required to deal with this problem. Smoothed models may be *interpolated* (where the probability estimate for a given N-gram depends on those for the corresponding (N-1)-grams) or *backed-off* (where the estimates for N-grams which occur at least once in the training data are determined whilst ignoring details of lower order (N-i)-gram statistics). Alternatively, they may use different forms of *discounting* (to adjust the probability estimates for N-grams which occurred fewer than (say) 5 times in the training data) and have different ways of estimating the probabilities of lower order M-grams (for M<N). A detailed discussion of such methods as linear interpolation between N-grams, smoothing methods, simulation of "unseen" events and use of "equivalence classes" can be found in Ney, Essen & Kneser (1994), Chen & Goodman (1999), Chen & Rosenfeld (2000) and Katz (1987).

Variants on the N-gram theme, such as variable sized N-grams (Siu & Ostendorf 2000), class-based N-grams (where classes of words - such as synonyms or parts of speech - are used instead of individual words in the N-grams) (Niesler & Woodland 1996) and "long distance" N-grams (where the next word is again predicted on the basis of (N-1) previous words, but where these words may be some distance back in the "history") (Rosenfeld 1996), have been tried by various authors. These have shown some improvement over simpler N-gram models, but still have major deficiencies.

Nevertheless, although there have been many attempts to improve on trigrams - and although such improvement is theoretically possible (Jelinek 1991) - most such attempts have met with very limited success (Jelinek 1991, Rosenfeld 1996, Brill et al 1998). The great success of trigram models is due to the facts that they are usually well-trained on the available data and that most modern European languages (which have been the focus of the majority of research on automatic speech recognition to

date) have a relatively strict word order and hence tend to have strong local dependencies between words (Jelinek 1991).

How can we improve on the performance of trigram models ? The following sub-sections review some of the methods which have been tried.

2.2 Adaptive (Dynamic) Models

The N-gram models (for N small) described previously are often considered as static models since they cannot use any information except the immediate history of the word sequence and are hence unable to adapt to the style or topic of the utterance. A logical way of improving on this would appear to be to allow a dynamic (or adaptive) model which can modify its probability estimates as a consequence of knowledge of the current conversation or monologue in order to improve its performance (Lau, Rosenfeld & Roukos, 1993). This should show benefits in two key areas : (1) if the overall conversation is quite diverse (or "heterogeneous"), but is made up of coherent ("homogeneous") chunks, each with a more consistent style, topic or vocabulary than the overall conversation. In this situation, an adaptive model should be able to focus on the properties of each homogeneous chunk and adjust its probability estimates accordingly. (2) It is not feasible to train a language model in a way which will be entirely suitable for every domain to which it could be applied. For example, a model trained on news stories would not be appropriate for use in a technical field or in an automated telephone switchboard. However, an adaptive model should be able to adjust to the language appropriate to the new application to which it is being applied (Lau et al 1993). Examples of adaptive language models are cache models (see below) and trigger based models. Whilst it is not claimed that either of these types of model actually reflects how speech recognition is carried out in the human brain, it is interesting to note that both of these have analogues in psycholinguistic models of human speech recognition : trigger models are somewhat like the psycholinguistic phenomenon called "priming" (Tillmann & Bigand 2003, Fischler & Bloom 1980, Meyer & Scvaneveldt 1971), whilst cache models have also been used in psycholinguistics (Walker 1998). Such analogies and relations will be further discussed below.

2.2.1 Trigger Pair Models

The concept of a trigger pair is in principle very simple. In a coherent document or conversation, certain words tend to be correlated with certain other words. For example, in a conversation or article about business and finance, the word "stocks" will often occur near the word "shares", as will various other words such as "price", "investment", "markets", "rose" and "fell". If the presence of one word (or sequence of words) A is strongly correlated with another word (or sequence) B, in the sense that the presence of A seems to raise the probability of also finding B present, then we say that "A triggers B" (A → B), with A being the "trigger" and B the "triggered word or sequence" (or "target"). In a similar manner, the fact that a word - particularly a relatively uncommon word - occurs once in a document or conversation frequently raises the probability of finding that word again later in the same text. For example, the presence of the word "investment" tends to suggest the current topic of interest is related to finance, and so there is a relatively high chance of finding the same word - which would normally be relatively uncommon - again. Such a situation, where A triggers itself (A → A), is called a "self-trigger" (Lau et al 1993, Rosenfeld, 1996).

As noted above, this statistical observation is analogous to the psycholinguistic phenomenon known as "priming" - where a person's "recognition performance" (in terms of response time, or success rate in a recognition or naming task) is found to be enhanced when the person has been primed by exposure to other words related to the target in meaning and/or sound (Bodner & Masson 2003, Meyer & Schvaneveldt 1971). Similarly, "inhibitory priming", where priming words are deliberately chosen to be unrelated to the target word, is generally found to degrade recognition performance.

Trigger pairs can be used in conjunction with a conventional model such as an N-gram by allowing the hybrid model to adapt probability estimates for proposed new words according to whether or not the new word has been "triggered" by an earlier part of the sequence - the probabilities of triggered words are increased, whereas those of non-triggered works are diminished. Trigger pairs provide a way of incorporating long-range relations between words into a language model.

In principle, even if the trigger and triggered objects are restricted to being single words, a system with a vocabulary of V distinct words will have V^2 theoretically-possible trigger pairs. However, the higher the number of trigger pairs used in the system, the more complex it becomes and it is therefore not feasible to incorporate all possible trigger pairs into the adaptive model - a choice must be made so that only the "most useful" triggers are used. Clearly, although the occurrence of one very common word may frequently be linked with another common word later in the text, this may not be very useful - for example "and" may often be followed at some later stage by "but" (although the actual correlation coefficient between these words may not be high). However, evidence based on a study (Rosenfeld 1996) of the Wall Street Journal corpus (Paul & Baker 1992) suggests that strongly-correlated pairs which occur very infrequently are also of limited value. For example, the presence of the proper noun "Brest" may be strongly correlated with the presence immediately afterwards of the proper noun "Litovsk", but such a link is unlikely to be a useful trigger pair - except in the contexts of discussions about World War I or Belorussian geography - since both words are very uncommon. On the other hand, a less-strongly correlated pair such as "stocks" and "shares" may be sufficiently common but yet sufficiently rare for ("stocks" → "shares") to be a worthwhile trigger to include (Rosenfeld 1996). A useful way to assess the likely utility of a trigger pair (A → B) as a means of using A to predict B is the pair's mutual information, I(A,B) :

$$I(A,B) = P(A,B) \log(P(B|A) / P(B)) + P(A,B') \log(P(B'|A) / P(B')) + P(A',B) \log(P(B|A') / P(B)) + P(A',B') \log(P(B'|A') / P(B'))$$

where, if A indicates that the appropriate word or sequence is present, then A' represents that it is absent, and the probabilities are estimated from statistics obtained from a large training corpus. Only those trigger pairs showing the highest mutual information should be included in the model.

Rosenfeld (1996) investigated a variety of trigger-type models : simple word triggers and "class based" triggers (i.e. based on groups of words which are related in some way, such as by meaning or by function), both *binary* triggers ("on" if the trigger is present in the "history" of the document or conversation, "off" if it is absent) and

frequency-dependent (based on how many times the trigger has occurred in the history) and distance-dependent triggers (taking account of how far back in the history the trigger last occurred). His general conclusions were :

(i) Different trigger pairs behave quite differently, and hence need to be modelled differently. The modelling should be more detailed when the expected benefits are greater.

(ii) Self-triggers are particularly powerful, to the extent that for over two-thirds of the words studied, the word itself was the trigger with the highest mutual information, and was one of the top 6 triggers for some 90% of words studied.

(iii) Triggers of the same linguistic root (e.g. "govern", "governor" and "government") are also generally powerful.

(iv) The majority of the best trigger pairs are associated with relatively common words, rather than word pairs which are strongly correlated but uncommon - see the "Brest" → "Litovsk" example above.

(v) Negative triggers - where the trigger and triggered word are from different topic areas (for example, the trigger being financial and the triggered word relating to agriculture), and hence likely to reduce the probability of the triggered word - can be of some, if limited value.

Rosenfeld (1996) did not make extensive use of the distance dependent triggers. In simple trigger models (in contrast to cache models - see below), the recent history of the document or conversation is treated as just a "bag of words" (Beeferman, Berger & Lafferty 1997). A word that appears just a few steps back in the history is regarded in exactly the same way as a word several hundred steps back. This does not seem entirely satisfactory. Some types of trigger (such as the "Brest" → "Litovsk" example) will be of most value at very short range (one step in this case), whereas others - often relating to the topic of discussion - will be of considerable value at greater distances. Some syntactic or stylistic-based triggers may also be useful at a longer range, due to the presence of embedded clauses, etc.. Beeferman et al (1997) have suggested that "attraction" (i.e. positive triggering) between words decays exponentially with distance, whereas stylistic and semantic constraints create a "repulsion" (negative triggering or inhibition) between closely-related words that discourages them occurring too close together. (The behaviour of self-triggers was found to differ slightly from that of other triggers - the probability of a given word

occurring was found to peak around 25 steps after the last occurrence of the same word.). They produced a three-parameter exponential model using a two-stage queuing process which incorporated both these phenomena, and found that a system incorporating this model showed improvement in performance over one using a simple trigger model in its place.

However, the assumption that only a relatively recent part of the history needs to be taken into consideration is not without justification. There is evidence from psycholinguistic studies of conversation that some information (particularly "surface" features such as syntax) is only retained in the short term by the participants (see Fletcher 1994 for a review of such work).

One drawback of trigger-based models is that they tend to be highly domain or topic sensitive. Furthermore, training such a model on a large amount of data from a corpus not in the area of its direct application (where perhaps only a very limited amount of data is available) may not be of much benefit (Rosenfeld 2000b). For example, use of the entire Wall Street Journal corpus (40 million words) or the entire Broadcast News Corpus (140 million words) gave little improvement to a system otherwise trained on a mere 2.5 million words of the Switchboard corpus (Godfrey, Holliman & McDaniel, 1992) - the system being intended for use with the type of conversational speech that the Switchboard corpus was based on - suggesting that current adaptation techniques are not sufficiently sophisticated (Rosenfeld 2000b).

A trigger-based model can be integrated with another model, such as an N-gram model , using an interpolation method or using the maximum entropy method (described in section 2.4 below) (Lau et al 1993).

2.2.2 Cache Models

Cache models (Kuhn & De Mori, 1990, Jelinek et al 1991, Clarkson & Robinson 1997, Iyer & Ostendorf 1999) have certain things in common with trigger models - they both use the *history* of the document or conversation to modify the probabilities of the candidates for the next word in the sequence. However, unlike trigger models, cache models tend to only consider a relatively small part of the most recent history (say the last 1000 words encountered). The simplest cache models, working on a "least recently used - out" principle - words which were included in the cache but have not occurred in the very recent history will eventually be replaced by others which have. The term *cache* is used by analogy with the concept of *cache memory* used in computer architecture and hardware. The cache is used to modify probabilities of the words it contains - based on their frequency of occurrence in the recent history rather than relying on the language model alone. However, although this simple form of cache can allow adaptation over a time period (measured in words) comparable to the size of the cache, it is unable to account for dependencies on a shorter scale such as those within a sentence (Iyer & Ostendorf 1999). Clarkson & Robinson (1997) have employed an exponentially decaying cache in an attempt to redress this, studying the hypothesis that even within the cache, the most recent words are those most likely to re-occur. Thus, their cache model assumes that a word's recurrence probability includes a contribution which decays exponentially with the distance between its last occurrence and the word of current interest. They compared the performance of simple caches of various sizes with those of caches with different "rates of decay" on data (both text and spoken) taken from the British National Corpus, finding that, at the optimal decay rate, their novel cache outperformed the best of the simple caches, which in turn performed significantly better than a basic trigram model. Some evidence from corpus studies (e.g. Purver, Ginzburg & Healey 2002) suggests that responses to utterances such as clarification requests in dialogue are much more likely to occur promptly after the request than some time later. Such findings would provide some justification for the use of decaying caches.

Walker (1996, 1998) has used a cache model (or "linear recency" model) to explain certain psycholinguistic phenomena relating to humans having limited attention whilst participating in discourse : utterances which are "informationally redundant,

difficulties in processing or recalling entities relevant to the discourse which are not "linearly recent" and experimental evidence that humans have limited attentional capacity (e.g. Miller 1956, Baddeley 1986, Fletcher 1994).

2.2.3 Cluster-Based Models

An alternative approach is to those discussed above is to create a separate language model for each class of a set of "characteristic classes" or "clusters" of texts. The motivation behind this type of approach is that it might be expected that texts might naturally be grouped in some manner (topic, word content, etc.) in such a way that the probability distributions of words varied significantly between the different groups. Thus, if it were possible to know, or decide in a reliable manner, which group, or cluster, a particular text belonged to, better results should be obtained using a language model constructed specially for that cluster than if an otherwise similar "general purpose" language model, derived from the contents of all the clusters, were used.

Note that in the context of the discussion here, we are primarily concerned with the clustering of sentences or dialogue turns. Other authors (e.g. Ney, Essen & Kneser 1994) have taken the approach of using clusters of individual words in order to tackle the problem of estimating probabilities of words which occur only rarely, or not at all, in the training data. This is to some extent similar to the approach used in our earliest experiments (Huckvale & Hunter 2001), but is not the issue of this present section.

According to Carter (1994a,b), use of clusters should enable a language model to encapsulate important contextual effects within sentences irrespective of how complex the probabilistic relationship between the relevant objects (words, word pairs, parts of speech, etc.) is. Furthermore, the clustering approach – unlike, say the use of N-grams for large values of N – does not make excessive new demands for training data. An additional benefit is that the clustering methodology can be applied in a standard way across different knowledge sources and for different kinds of documents. If the language models based on clusters do turn out to be better than the corresponding models not using clusters, then there is good evidence that there are connections between the objects comprising the clusters which the original (not cluster-based) model fails to exploit.

Presented with a new document or utterance, an assessment could be made as to which class this document should be assigned, then the language model for that class applied to it. Of course, the initial judgement of class for the new document may not be correct, particularly if (as is likely to be the case in practice) the decision is made on the basis of incomplete information about the new document. However, this decision need not be a permanent one, or different strategies for selecting the most appropriate cluster can be investigated. Alternatively, the new document could be specified as being "quite a lot like cluster A, a bit like cluster B, but not so much like cluster C" – a weighted mixture of clusters. In principle, the clusters could be assigned by human decision (e.g. manually "marking-up" or "tagging" the training data), but this is highly labour-intensive. Thus, it is commonly preferred to assign documents to clusters by an automatic process. The criteria used for initially identifying clusters could be topic-based, or based on similarities (in terms of their lexical content) between the documents comprising each cluster (Sekine 1994, Robertson & Spärck Jones 1997, Iyer & Ostendorf 1999). Such methods have some appeal from a linguistic perspective. Alternatively, "engineering" approaches have been developed which construct clusters iteratively in a manner which will optimise the perplexity (as defined in section 1.3) of a language model based on the current cluster with respect to the current text (Carter 1994 a, b, Clarkson & Robinson 1997). Such approaches lack the linguistic motivation of those mentioned above, but may result in equally good, or even better, performance.

2.2.3.1 Linguistically (Lexically)-Motivated Clustering Methods

Such methods cluster sentences of dialogue turns according to their lexical content. Broadly speaking, sentences (or turns) containing largely the same distinct words will be put in the same cluster. This approach is based on how "characteristic" of, or "distinctive" to, each cluster of documents any particular word is. This is essentially the approach employed by Sekine (1994) and Iyer & Ostendorf (1999), who define a similarity metric S_{ij} between documents (or clusters) i and j by

$$S_{ij} = \frac{1}{|A_i| . |A_j|} \sum_{w \in A_i \cap A_j} \frac{N_{ij}}{|B^{(w)}|}$$

where A_i is the set of distinct words in document i (and hence $|A_i|$ is the size of that set), $B^{(w)}$ is the set of documents which contain the word labelled w and N_{ij} is a normalising factor :

$$N_{ij} = \sqrt{\frac{N_i + N_j}{N_i . N_j}}$$

with N_i being the number of documents in cluster i. The normalising factor was introduced in order that, when applying the metric to construct clusters, the clustering process countered the tendency for small clusters to join large clusters, rather than several small clusters joining together (Iyer & Ostendorf 1999). This type of metric is called a "combination of inverse document frequencies" and one of its positive features is that words which are generally very common (i.e. common in very many documents) – such as function words – have little effect on the clustering process. Note that this metric gives a larger result when the two clusters are very similar ($S_{ii} = \sqrt{(2 / N_i)} / |A_i|$ if each word in A_i only occurs in A_i and no other cluster), but S_{ij} is zero if clusters A_i and A_j contain no words in common.

Robertson & Spärck Jones (1997) have proposed and used a rather more sophisticated metric, but with a similar motivation. Their metric has three components :
(i) "collection frequency weights" (similar to the "inverse document frequencies" of Sekine (1994) and Iyer & Ostendorf (1999)), which note that words which only occur in a relatively small number of documents are usually more valuable in judging document similarity than words which occur in very many documents. Thus, the appearance of the word "Litovsk" in two distinct documents would suggest that those documents have a high chance of being about a similar theme, or at least have several other words in common.
(ii) "term frequencies", the frequency of a word's occurrence within a given document – the more frequently a word (or at least a word which is not normally very common) occurs in a given document, the more relevant it is likely to be for that particular

document. For example, if the word "shares" occurs in a document much more often than we would routinely expect, then it is quite likely that the document has a financial theme and we might expect other documents with this theme to appear in the same cluster.

(iii) document length – a given word is more likely to occur several times in a long document than in a short document (all other factors, such as document topic, being equal). Thus, a word which occurs a large number of times in a relatively short document is more likely to be important in categorising that document than if the same word had occurred the same number of time in a longer document.

Robertson & Spärk Jones (1997) combine these factors in various ways to give several measures for a word (or "term") t(i) labelled i in a document d(j) labelled j. If n_i is the number of documents in which the term t(i) appears, and N is the total number of documents which we are considering, then the "collection frequency weight" for that term is :

$$CFW(i) = \log(N) - \log(n_i) = \log(N / n_i)$$

The "term frequency", TF(i,j), is simply the number of times term t(i) occurs in document d(j), and the "document length", DL(j), is just the total number of term (word) occurrences in document d(j).. However, it is often more practical to work with the "normalised document length", NDL(j), found simply by dividing DL(j) by the mean value of DL across all documents being considered.

They propose the "Combined Weight" for term (word) t(i) and document d(j) as :

$$CW(i,j) = \frac{CFW(i) \; . TF(i,j) \, . (1 + K_1)}{\left[TF(i,j) \;\; + \;\; K_1 \, . ((1-b) + b.NDL(j)) \right]}$$

where K_1 and b are adjustable parameters. K_1 adjusts the influence of the turn frequency – setting K_1 to zero removes any influence of TF on CW, whereas the higher the value of K_1, the stronger the influence of TF. Robertson & Spärck Jones (1997) reported that use of $K_1 = 2$ had proved to be effective in tests on text documents such as news items and government reports. The parameter b, restricted to values between 0 and 1, adjusts the influence of both document length and term frequency on CW. If b = 0, then it is being assumed that documents tend to be long if

they contain many distinct topics, whereas if b =1, the length of documents is assumed to be due to repetitiveness and verbosity. Robertson & Spärck Jones (1997) suggested that values of b around 0.75 had been found to work well in their studies.

Robertson & Spärck Jones (1997) propose for further developments to this similarity measure, particularly for use in iterative schemes where, following some initial assessment, documents can be marked as "relevant" or "irrelevant" and the (term, document) weights re-evaluated. One such approach is the use of "relevance weights". These take note of the observation that a word may be present in many relevant documents just because it is present in a lot of documents – relevant or not. Essentially, the relevance weight is a modified form of the collection frequency weight described above. If r_i is the number of documents marked as "relevant" which contain the term (word) t(i) and R is the total number of documents marked as "relevant", then an estimate of the relevance weight is given by :

$$RW(i) = \log\left(\frac{(r_i + 0.5).(N - n_i - R + r_i + 0.5)}{(n_i - r_i + 0.5).(R - r_i + 0.5)}\right)$$

where N and n_i have the same meaning as before. This is only an estimate of the true relevance weight, since we should not assume that a term will never be found in any relevant document just because, on the basis of a relatively small sample, it has not been seen in any relevant documents to date (Robertson & Spärck Jones 1997). This observation, and the need to avoid "singular" arguments for the logarithm function which would occur if either numerator or denominator were to become zero, explains the use of the 0.5 added in each term of the formula above as a correction to remove bias due to a finite sample size.

The relevance weight can then be incorporated into a revised form of the combined weight, called the " combined iterative weight", CIW :

$$CIW(i,j) = \frac{RW(i) \;.TF(i,j).(1+K_1)}{\left[TF(i,j) \;\; + \;\; K_1 \;.(\,(1-b) + b.NDL(j)\,)\right]}$$

The metric chosen for use can be applied in conjunction with a k-means clustering algorithm to form clusters :

Suppose we have N documents to be grouped in k clusters, where $N >> k$:

(1) Randomise the order of the documents.
(2) Assign the first k documents in the randomised list to clusters 1 to k respectively, i.e. if $1 \leq i \leq k$, then document i is put in cluster i.
(3) For each document, using the chosen metric, check whether there is any cluster which has a mean point closer to the current document than the mean point of the cluster to which that document is currently assigned.
(4) Repeat (3) until fewer than a specified number, T, of documents change clusters during one iteration.

2.2.3.2 Clustering Methods Based on Perplexity (or Entropy)

Carter (1994) and Clarkson & Robinson (1997) instead chose to base their similarity measure on the quantity which, for the language model whose construction was the eventual aim, they were trying to optimise : the perplexity of a language model with respect to text from the domain(s) of interest. In the ideal case, the measure would be the perplexity of the language model, preferably a trigram (or equally sophisticated) model, which would result if a specified cluster (in its current state) were used for training data and the document of current interest as test data. However, this would require an infeasibily high amount of computation to be carried out at each step of the iterative clustering process. Instead, the metric used is the perplexity of a unigram language model : the "distance" between a given test document and a cluster is the perplexity of the unigram model trained on all the text in that cluster with respect to the text in the test document (Clarkson 1999, p52). If N_C and N_D are the total number of words in the current cluster C and the test document D respectively, $N_C(w)$ and $N_D(w)$ are respectively the number of time that the word w occurs in the cluster and the test document, and $\underline{w}^{(D)} = (w_1, w_2, \ldots, w_{Nc})$ is the test document written as a word string, then from the definition of perplexity, the required unigram model perplexity can be calculated as :

$$PP_1 = \left(P(\underline{w}^{(D)} \mid C) \right)^{-1/N_D}$$

$$= \left[\prod_{i=1}^{N_D} P(w_i \mid C) \right]^{-1/N_D} \text{ by the assumed independence of unigram probabilities.}$$

Naively, we could estimate $P(w_i \mid C)$ by $(N_C(w_i) / N_C)$, but this would assign a zero probability to all words which had not already been observed in the cluster C. Instead, a "floor value", δ, can be added to the counts to ensure that ensure non-zero counts in the cluster for each word present in the text (whether or not it is genuinely present in the cluster), and hence guarantee finite perplexity values :

$$PP_1 \approx \left[\prod_{i=1}^{N_D} \left(\frac{N_C(w_i) + \delta}{N_C} \right) \right]^{-1/N_D} = \frac{1}{N_C} \left[\prod_{i=1}^{N_D} (N_C(w_i) + \delta) \right]^{-1/N_D}$$

Note that, unlike the lexically-motivated metrics discussed in the previous section, this perplexity metric has a relatively low value for similar documents or clusters and higher values for dissimilar ones. Although this metric is less *explicitly* based on the particular words within a document cluster, it is still *implicitly* highly dependent on them – since the relevant entropy or perplexity values are dependent on word probabilities given by the language models for these clusters.

Once the metric has been defined, clusters can be constructed using a k-means algorithm, as described in section 2.2.3.1 above. Once the contents of all the clusters have been finally decided upon, a separate trigram language model can be constructed for the documents in each cluster. In application, some criterion must be used to decide which cluster-based model (or combination of models) would be most appropriate for use with the utterance or document currently under consideration – this will be discussed in more detail in Chapter 7, where the cluster-based experiments will be described.

2.3 Combining Information from Several Sources

2.3.1 Introduction

When we wish to combine probabilistic information from several sources – for example, acoustic information from an acoustic model and linguistic information from a language model, or from several different types of language model (e.g. a trigram model and a cache model), we require a methodology for combining probabilities from the different sources in order to obtain a valid probability as the result. There may also be good reason to give more "weight" to values from certain sources than others.

There are two obvious approaches to combine probabilities from distinct sources (sometimes called different "probability streams") : in a multiplicative manner or in a weighted averaging manner.

Consider a relatively general case where we are trying to predict the probability of an event x based on probabilistic information from M distinct sources (streams), which we assume to be independent. Let P(x) be our overall estimate of the probability of x occurring and $p_i(x)$ be the probability of x occurring according to stream i (for $1 \leq i \leq M$). In most cases of interest here, the probabilities will all be conditional on some specified history or events having occurred.

In the multiplicative approach,

$$P(x) = \prod_{i=1}^{M} K_i \left(p_i(x)\right)^{\alpha_i}$$

where K_i is a normalising factor (introduced to ensure $\sum_{all\ x} P(x) = 1$) and α_i is an exponent to account for the relative importance of the different information sources. This will require selection of appropriate values of the K_i and α_i . The maximum entropy approach (see section 2.4 below) is a suitable method for selecting these parameters in a way which is consistent with a body of training data.

The weighted-averaging approach is normally known as linear interpolation, and is described in section 2.3.2 below.

2.3.2 Linear Interpolation

Linear interpolation is an "additive" or "weighted averaging" approach to conbining information from several sources. Using the same notation as in section 2.3.1 above, the overall estimate of the probability of event x occurring, based on the M information sources is given by :

$P(x) = \sum_{i=1}^{M} \beta_i \, p_i(x)$, where the β_i are weighting factors, to be determined, showing the relative importance of the information sources, such that $0 \leq \beta_i \leq 1$ for each i and $\sum_{i=1}^{M} \beta_i = 1$. There are therefore (M-1) parameters to be selected (the final one is then determined by the others). In the context of statistical language modelling, a method commonly used for selecting these weights is to use the "Expectation-Maximisation" (EM) algorithm (Dempster et al 1977, Jelinek 1990) which chooses the weights in a way which optimises the perplexity of the resultant interpolated model with respect to the data used for the purpose. More explicitly, the EM algorithm is a method originally intended for estimating the optimal set of parameters $\{ \beta_i \}$ which maximise a likelihood function $g (\underline{y} \mid \{ \beta_i \})$ for a given set of observable data $\underline{y}$ dependent on non-observable values $\underline{x}$ which are distributed with probability density $f (\underline{x} \mid \{ \beta_i \})$. The EM algorithm is a two-phase iterative process for estimating these optimal parameters :

Phase 1 (E Phase):
Find the Expectation $E(\log (f(\underline{x} \mid \{ \alpha_i \}) \mid \underline{y}, \{ \beta_i \})) = Q(\{ \alpha_i \} \mid \{ \beta_i \})$, say, where there are the same number of parameters $\{ \alpha_i \}$ as $\{ \beta_i \}$. We assume that this function exists for all possible pairs $(\{ \alpha_i \} , \{ \beta_i \})$ and that $f(\underline{x} \mid \{ \beta_i \}) > 0$ "almost everywhere" in the space of all possible $\{ \beta_i \}$.

Phase 2 (M Phase): "Maximise" (optimise) $Q(\{ \alpha_i \} | \{ \beta_i \})$ over the possible values of $\{ \alpha_i \}$. Choose $\{\alpha_i^* \}$, the set of values of the $\{ \alpha_i \}$ which give this maximum in Q, to be the new values of the $\{ \beta_i \}$.

Repeat phases 1 and 2 in turn until some convergence criterion is met.
Ideally, it would be preferable to maximise $\log (f (\underline{x} | \{ \beta_i \}))$, but this quantityis not known explicitly, so we have to use the best estimate of it currently available : its expectation given the data $\underline{y}$ and the current best estimates of the parameters $\{ \beta_i \}$ (Dempster, Laird & Rubin 1977).

This algorithm will eventually converge to the values of the parameters which are optimal with respect to the data used for their calculation. If this data set is sufficiently large and is representative of the test data (to which the resulting model is to be applied), then the parameters should also be close to optimal with respect to the test data (Rosenfeld 1994).

The advantages and disadvantages of using linear interpolation as a means of combining information from different statistical language models have been discussed by Rosenfeld (1994). To summarise his findings :

Advantages :
(i) Linear interpolation is very general, and can be used with all kinds of language models. In fact, only the probability streams, not the models themselves, are needed in the actual interpolation process.
(ii) Linear interpolation is simple to implement, experiment with and analyse. Packages, such as the Cambridge-Carnegie Mellon University Language Modelling Toolkit (Clarkson & Rosenfeld 1997), often include programs to perform linear interpolation of models. (Indeed, it is the *interp* program from this package which has been used for this purpose in the experiments described later in this thesis.) The weights do not normally need to be specified very precisely (changes of up to about 5% make little difference to the perplexity of the interpolated model) and relatively little data (several thousand words per parameter) needs to be held back for training the weights.

(iii) A linearly interpolated model cannot be worse than any of its individual components, at least with respect to the data on which is weights are trained. Normally, if an additional source of information is of little or no use to a composite language model, the EM algorithm will result in that component having a very small weight in the interpolated model.

Disadvantages :

(i) Linear interpolation does not always make optimal use of the available information sources. The different components are consulted "blindly" without regard to their individual strengths and weaknesses, which may be context-dependent, and the weights are optimised globally (over all the data reserved for that purpose), rather than with respect to data of current interest. This will make the resulting model less adaptable. Furthermore, the interpolated model is a weighted arithmetic average, whereas perplexity, the commonly-used measure of language model "quality", is a geometric average, so small-weighted contributions to linearly interpolated models can make large reductions in perplexity compared with the original language models.
(ii) Linearly interpolated models can be statistically inconsistent with their individual components. The different component models will probably partition the data space in different ways, and an interpolated model (with weights computed globally over all the data reserved for that purpose) will not be able to produce probabilities which are entirely consistent with counts of observations for any one such partition.

Nevertheless, in spite of its disadvantages, linear interpolation is still the most commonly used means of combining information from distinct sources in statistical language modelling, and will be extensively used for this purpose in the remainder of this study.

2.4 The Maximum Entropy Method

2.4.1 Introduction

The idea of maximum entropy inference has its origins in the work of Boltzmann (c. 1870) and Gibbs (c. 1900) in statistical physics and thermodynamics (Jaynes, 1988). However, in some way it has similarities with Occam's Razor - the principle due to

the Medieval philosopher William of Occam, which suggested that in situations where two explanations were offered for a phenomenon, in the absence of additional information, the simpler explanation should be preferred (Berger at al 1996, Gull 1988) - and Laplace's "Principle of Insufficient Reason" (1843) (Jaynes 1988).

An informal statement of the principle of maximum entropy can be stated as : "In order to produce a model which is statistically consistent with the observed results, model all that is known and assume nothing about that which is unknown. Given a collection of facts or observations, choose a model which is consistent with all these facts and observations, but otherwise make the model as 'uniform' as possible." (Berger et al, 1996).

As an example of this, consider the problem of reducing the task of translating from English into French to a set of statistical rules which could be implemented on a computer (Berger et al 1996). There are several French words and phrases which may be considered as translations of the English word "in" - "dans", "en", "à", "au cours de" and "pendant" are probably the most common. According to the principle of maximum entropy, *in the absence of further information*, we should choose a model which makes a uniform choice between these whenever the English word "in" is encountered, i.e. p(dans) = p(en) = p(à) = p(au cours de) = p(pendant) = 1/5 . Clearly, this model does not include any sophisticated knowledge about the French language, but is purely the most uniform model in the absence of further information - as directed by the principle of maximum entropy.

Now suppose that, on consulting an expert, we found that 30% of the time the appropriate translation for "in" was either "dans" or "en". We can now revise our model in the light of this observation. Given this, the five options above, but no extra information, the "most uniform model" we can choose would be :
p(dans) = p(en) = 3/20 but p(à) = p(au cours de) = p(pendant) = 7/30
We have chosen the "most uniform" probability distribution consistent with the observation that p(dans) + p(en) = 3/10 (since "dans" and "en" account for 30% of the occurrences of "in"), and for the remaining options we have chosen the most uniform distribution consistent with the requirement that

$$p(\text{dans}) + p(\text{en}) + p(\text{à}) + p(\text{au cours de}) + p(\text{pendant}) = 1.$$

If we had further observations, such as 'The expert translates "in" as either "dans" or "à" 50% of the time', then it may become less obvious how to modify the model whilst maintaining consistency with the data and keeping the model "as uniform as possible". The maximum entropy method provides a framework for achieving these objectives.

This still contains very little knowledge of the French language. Further refinements could be made by adding context-sensitive data, for example 'If "in" follows the word "April", in 90% of cases, the French translation is given as "en" by our expert.' The probabilities in our model now become conditional probabilities.

2.4.2 Mathematical Framework : Feature-Based Models and Maximum Entropy

The analysis below follows that of Berger et al (1996).

Consider a set of data, in which the response of the system in which we are interested is observed as y under conditions (or context) $\underline{x}$. In the translation example above, y would be the French translation and $\underline{x}$ would be the context of the English word being translated - including the word itself, the remainder of the sentence in which it appears and possibly additional information such as the topic of the text or speech. We can then collect a set of data where we have translations $\{y_i\}$ corresponding to contexts $\{\underline{x}_i\}$, giving a set of ordered pairs $\{(\underline{x}_i , y_i)\} = \{(\underline{x}_1 , y_1)\} , (\underline{x}_2 , y_2) , \ldots , (\underline{x}_N , y_N)\}$. In our translation example, these would be the translations, y_i , of "in" offered by a expert for various English sentences, $\underline{x}_i$. Within a statistical model we can use $P(y \mid \underline{x})$ to denote the probability that y is the output given input context $\underline{x}$, or the conditional probability of y occurring given $\underline{x}$.

Given a sample of N items of training data (data to be used to train the model), we can denote the *empirical* bivariate probability distribution of $\underline{x}$ and y co-occuring by $\mathbb{p}(\underline{x} , y)$:

$\mathbb{p}(\underline{x} , y)$ = (Number of times that $(\underline{x} , y)$ occurs in the training data) / N

Note, however, that there is only likely to be a few occurrences of any given ($\underline{x}$, y) pair in the training data set, and many possible pairs may not occur at all. This can make reliable estimation of probabilities from limited data quite difficult. (Note that throughout the discussion below, the symbol $\mathbb{p}$ will refer to an empirically-estimated probability distribution, whereas p will refer to the true underlying distribution.)

We now wish to construct a statistical model which will enable us to predict the most appropriate y for a given context $\underline{x}$, whilst being consistent with the training data. We will make use of Bayes' Theorem of Conditional Probability :

$$P(A, B) = P(A|B).P(B) = P(B|A).P(A)$$

and hence inference techniques of this type are sometimes known as Bayesian methods.

We can reduce each the context-outcome pair ($\underline{x}$, y) to a vector of binary features $\{f_i\}$. Each feature is "on" or "off" (1 or 0) depending on the ($\underline{x}$, y) pair currently under consideration. For example, in the English to French translation problem, f_1 could be set to 1 if and only if y is "en" and the English context $\underline{x}$ has "in" followed by "April".

Noting that $\mathbb{p}(f_i) = \mathbb{p}(f_i = 1)$, the expected value of a feature f_i with respect to the empirical (observed) probability distribution of contexts and outcomes is given by :

$$E_e(f_i) = \mathbb{p}(f_i) = \sum_{\underline{x}, y} \mathbb{p}(\underline{x}, y)\, f_i(\underline{x}, y)$$

whereas the expected value of f_i with respect to the conditional probability model we are constructing is :

$$E_m(f_i) = p(f_i) = \sum_{\underline{x}, y} \mathbb{p}(\underline{x})\, p(y \mid \underline{x})\, f_i(\underline{x}, y)$$

where $\mathbb{p}(\underline{x})$ is the observed probability distribution of contexts $\underline{x}$ in the training data.

We require that the model is consistent with the training data, and thus that $p(f_i) = \mathbb{p}(f_i)$. Hence, we obtain :

$$\sum_{\underline{x}, y} \mathbb{p}(\underline{x})\, p(y \mid \underline{x})\, f_i(\underline{x}, y) = \sum_{\underline{x}, y} \mathbb{p}(\underline{x}, y)\, f_i(\underline{x}, y)$$

as a constraint on our conditional probability model $p(y \mid \underline{x})$.

We now wish to make our model as "uniform" as possible, but how do we quantify "uniformity" ? The measure generally used is the entropy, H(p), of the probability distribution p :

$$H(p) \equiv -\sum_{i} p_i \log(p_i)$$

where the summation runs over all possible states in which the system may be found, and p_i represents the probability of the system being found in that state. Strictly speaking, the logarithm should be taken to base 2, but consistent use of a logarithm to any other base will just result in a scaling of the entropy by a constant factor.

In the context of the conditional probability distributions of interest here, we have :

$$H(p) = -\sum_{\underline{x}, y} p(\underline{x}, y) \log(p(\underline{x}, y)) = -\sum_{\underline{x}, y} \mathbb{p}(\underline{x})\, p(y \mid \underline{x}) \log(p(y \mid \underline{x}))$$

where the summations run over the complete set of possibilities for $\underline{x}$ and y.

A probability model with no uncertainty at all ($p(y \mid \underline{x}) = 1$ for all $\underline{x}$ and y) would give an entropy of zero, which is the lowest possible value of H(p). ($0 \leq p \leq 1$, so log(p) is always non-positive.) Similarly, a completely uniform model where $p(y \mid \underline{x}) = 1/|Y|$, with $|Y|$ the number of different possible values for y, independent of the context $\underline{x}$, will have entropy $\log(|Y|)$, which is an upper bound for H(p). Thus, the "more uniform" the model, the larger the value of H(p). This leads us to a more formal, mathematical statement of the principle of maximum entropy :

"In order to choose a probability model p from a set of permitted models, choose whichever model is which is consistent with the given data whilst maximising the entropy, H(p), of the model."

Thus, the problem of finding the "best" probability model p becomes a problem of optimisation under constraints :

Maximise H(p) with respect to p, whilst satisfying the constraints :

$p(f_i) = \mathbb{p}(f_i)$ for i = 1, 2, 3, ...

This can be tackled using the method of Lagrange multipliers. We convert the problem into one of unconstrained maximisation of the Lagrangian function :

$$L(p, \underline{\lambda}) = H(p) + \sum_i \lambda_i (p(f_i) - \mathbb{p}(f_i))$$

When the constraints of the original problem are satisfied, $p(f_i) - \mathbb{p}(f_i) = 0$ and so $L(p, \underline{\lambda}) = H(p)$.

We then find the maximum (with respect to varying p) of $L(p, \underline{\lambda})$ whilst holding $\underline{\lambda}$ fixed. Use p_λ to denote the value of p which achieves this for the current $\underline{\lambda}$ and $\Psi(\underline{\lambda})$ for the value of $L(p, \underline{\lambda})$ there.

Defining $Z_\lambda(\underline{x}) = \sum_y \exp (\sum_i \lambda_i f_i(\underline{x}, y))$

it can be shown using multivariate calculus that the optimal values, for a given $\underline{\lambda}$ are :

$$p_\lambda (y \mid \underline{x}) = (\exp (\sum_i \lambda_i f_i(\underline{x}, y))) / Z_\lambda(\underline{x})$$

$$\Psi(\underline{\lambda}) = - \sum_{\underline{x}} \mathbb{p}(\underline{x}) \log(Z_\lambda(\underline{x})) + \sum_i \lambda_i \mathbb{p}(f_i)$$

We then find the optimal value of $\underline{\lambda}$, $\underline{\lambda}^*$, which maximises $\Psi(\underline{\lambda})$. Substituting this into the expression for $p_\lambda (y \mid \underline{x})$ above will yield the most uniform model, or "model with maximum entropy", consistent with the data. It can be shown that, under appropriate (Kuhn-Tucker) conditions (Greig, 1980) , any algorithm which finds the value $\underline{\lambda}^*$ which maximises $\Psi(\underline{\lambda})$ will find the correct value of $p_\lambda (y \mid \underline{x})$ which maximises H(p). In the earliest experiments we performed (Huckvale & Hunter 2001), we chose to use the "downhill simplex method" (Nelder & Meade 1965, Press et al.

1992) because of its simplicity and reliability. However, this method proved to be somewhat inefficient in the number of steps – and hence the computational time – it took to find the optimum set $\underline{\lambda}^*$ where the set possible features is large. In more recent experiments, we have employed the "generalised (or improved) iterative scaling method" (Darroch & Ratcliffe 1972, Rosenfeld 1994, Berger et al 1996) which, although somewhat more complicated, proves to more efficient.

It can also be shown that the constrained optimisation problem of finding of finding the probability distribution with maximum entropy from the family of distributions $p_\lambda(y \mid \underline{x})$ is the dual problem to the unconstrained optimisation task of using maximum likelihood to find the set of parameters $\underline{\lambda}$ which maximises $\Psi(\underline{\lambda})$ (Berger et al 1996).

The above analysis does not explain how the features to be incorporated into the model should be chosen, or how many features should be incorporated into the model. Strictly speaking, these are not the concerns of the maximum entropy method. However, they are of great practical importance when solving real problems. It is best to start with a large selection of possible "candidate" features which can be any parameters which help describe the context. Features can then be added to the model in an iterative manner - adding the features one at a time, according to which feature seems to make the best improvement to the model at that stage and continuing until adding further features seems to make negligible improvement to the model. However, constructing models incorporating a large number of features from an even larger set of candidate features using maximum entropy is highly computationally intensive. It may be wiser to reduce the size of the set of candidate features by initially pre-selecting or "winnowing" using some measure of "probable usefulness" of each feature, such as mutual information between the feature and output value y (Rosenfeld, 1996).

The iterative process of feature selection is carried out by initially choosing a model containing no features. Some feature is then temporarily added to the model, and the resulting improvement to the model (in terms of gain in likelihood with respect to the training data), if any, measured. The feature is then removed from the model and another tried instead. This is repeated until all features have been tried, and the one which improves the model the most added permanently. The process is repeated

trying all candidates for a second included feature, then for a third, and so on. The process is terminated when the addition of further features makes negligible improvement to the model, or when the process becomes computationally infeasible. (The computational time required to construct a model containing N features from a fixed set of M candidates, with M significantly greater than N, seems to grow approximately exponentially with N.)

This process does not specify what sort of features are likely to be useful. Indeed, this will depend heavily on the precise nature of the problem being studied. In an area such as statistical language modelling, features could be any one of a wide range of things - lexical or topic information, pragmatic information, grammatical information to name but a few. The best policy is probably to err on the side of allowing a large number of possible features initially, then reduce the number being considered using a "winnowing" process, as described above.

One of the first major applications of the maximum entropy method was in astronomy for the purpose of reconstructing images from noisy data (see, e.g. Burch et al, 1983, Gull & Skilling 1984). However, the first applications of it to statistical language modelling seem to have been during the mid-1990's (Berger et al 1996, Rosenfeld 1996).

For an example of an experiment using Maximum Entropy to train an exponential probability model (using sequences of "word classes" based on parts of speech) which is then applied to a language modelling problem , see Huckvale & Hunter (2001) – in Appendix B of this thesis.

Experiments using models based on word trigger pairs, trained using Maximum Entropy, will be described in chapter 6.\

Chapter 3 Dialogue and Discourse

3.1 Motivation : What's Special About Dialogue ?

The main focus of this thesis is the application of statistical language modelling techniques to dialogue situations – in particular, the large body of transcribed British English dialogue material within the British National Corpus (see chapter 4). As will be discussed below, dialogue differs significantly from written text material both in its structure and lexical content. As noted by Taylor et al (1998), most automatic speech recognition systems have been designed to deal with read speech or isolated utterances and few systems have been adapted to take account of the differences between these types of speech and spontaneous conversational speech. From the perspective of speech technology, the study of dialogue is important for at least three reasons. Firstly, it would be hoped that incorporating features taking account of the nature of dialogue into automated spoken dialogue systems would lead to better word recognition (or utterance comprehension) performance. Secondly, such features could also aid more natural human-machine interaction. Thirdly, evidence from a large corpus of dialogue data could lead to a better theoretical understanding of, and empirical justification for, linguistic theories of dialogue, some of which will be briefly discussed below.

Thus, I believe that taking account of its special nature, rather than treating it in exactly the same way as ordinary text, will be beneficial in statistical language modelling. Previous authors (e.g. Rosenfeld 2000b) have noted that the success of statistical language models is very sensitive to the nature of the material on which they are trained and to which they are applied. However, to date, the majority of language models are based on either ordinary text or transcribed broadcast news (largely monologue) data. In recent years, relatively large corpora of transcribed dialogues have become available (see examples in table 3.1 below), which make serious statistical language modelling of dialogue data feasible.

Corpus	English	Signal	Speakers	Size	Style
SWITCHBOARD (1.2)	American	Telephone	543	~ 3 million words (~ 240 hours)	Given topic
CALLHOME-English	North American	Telephone	120	230 000 words	Conversation
CALLFRIEND-English	American	Telephone	60	~ 1200 minutes	Conversation
BRAMSHILL	British	Microphone	~200	600 000 words	Given topic
HCRC Map Task	British	Microphone	64	~ 18 hours (~1080 minutes)	Given topic
DCIEM Map Task*	Canadian	Microphone	~40	175 000 words (216 dialogues)	Given topic
British National Corpus (BNC)	British	N/A	> 124	7.7 million words	Conversation

Table 3.1: Some corpora of transcribed dialogues in English

Although several statistical studies have been carried out on some of the corpora listed in table 3.1, such as the SWITCHBOARD corpus (Godfrey et al 1992, Godfrey & Holliman 1997, Chelba & Jelinek 1999, Jurafsky et al 1998), most of these have either been in American English or contain only dialogues relating to a specific task, such as the HCRC Map Task corpus (Anderson et al 1991, Carletta et al 1997, HCRC 2001)*. Clarkson & Robinson (1997) and Clarkson (1999) have performed statistical language modelling studies on British English data, but using the entire BNC, without distinguishing between text, spoken monologue and spoken dialogue material. There would therefore appear to be scope for a study on statistical language modelling of British English dialogue data, such as that within the BNC, and that is the focus of this thesis.

*Most of the statistical modelling (Taylor et al 1998, King 1998, Wright et al 1999, Wright 2000) carried out on the Map Task was actually performed using the DCIEM Canadian English Map Task Corpus (Bard et al 1995, 1996) rather than the HCRC British English version. Note that much of the DCIEM corpus was acquired whilst the participants were suffering from some degree of sleep deprivation.

One of the aims of the project which resulted in the BNC was that it should be a large corpus, representative of modern British English – including examples of material from different regions of the UK, from speakers of a variety of ages and of different social groups. This would appear to make the dialogue material in the BNC particularly suitable for training a language model which could be applicable for use in a general dialogue system usable by any speaker of modern British English.

Spontaneous speech – even once transcribed - is rather different to written text. Disfluencies and speech repairs will tend to be common. In informal situations, colloquial expressions and slang are more common than in text, whereas complex words – technical words and "sophisticated" words of Latin or Greek origin – may tend to be more infrequent in speech than in text. Transcribed dialogue differs even further than transcribed monologue does from ordinary written text. The observation that the participants tend to take turns to speak, with the speaker normally changing at particularly appropriate points in the conversation, is unique to dialogue (and the multi-speaker equivalent, "polylogue"). Furthermore, there are issues of what knowledge, both explicitly related to the topic of the conversation and to more linguistic issues such as the referents of pronouns and other anaphora, is already shared between the participants and what needs to be negotiated between them. Co-operation between the speakers is clearly a requirement for a "successful" dialogue. Greetings and other phatic utterances have a definite role in dialogue, but are largely irrelevant in written text, if they occur at all. Structures between turns, such as question/answer pairs or question/clarification request/clarification/answer, can be important, and the way in which topics are introduced, developed and changed may be very different in dialogue from text or spoken monologue.

It would therefore be expected that incorporating some of these distinctive features of dialogue into automatic dialogue system – either within the language model, or as a separate module – should be beneficial to the performance of the system. Some previous approaches at attempting this will be discussed in the remainder of this chapter, whilst my own experiments applying statistical language modelling

techniques to the dialogue material within the British National Corpus (BNC) forms the theme for the remainder of the thesis.

3.2"Dialogue" and "Conversation" – Some Perspectives from Linguistic Theory and Psycholinguistics

Prior to discussing approaches to modelling dialogue (particularly from a computational or statistical perspective), it may be instructive to consider some concepts from linguistic theory relating to "conversation", "discourse" and "dialogue", and their implications for the computational modelling of dialogue. Some of these concepts can be incorporated into a computational or statistical model in a very straightforward way, whereas other features may be very difficult, or even impossible, to model.

Cameron (2001, pp 7-18) has discussed the distinctions between "conversation", "talk" and "spoken discourse" in some detail. Some of the material below follows the argument of her analysis.

"Conversation" is normally taken to mean spoken (as opposed to written) language. However, some would question whether the situation is that simple. Is monologue (speech by a single person) really "conversation" ? Are the utterances of a school or college class lesson, or those of an interview or a medical consultation, really "conversation" ? Perhaps "interactivity" (the speaker changing frequently, i.e. the majority of the individual speaker turns being relatively short) and spontaneity (the contributions of each participant are not the result of significant planning, unlike the case of, say, the teacher's role in the verbal exchanges in a school lesson) are necessary features of "true conversation" (Nofsinger 1991, pp3-4) ? Perhaps "conversation" is closer in meaning to "chat" or "gossip" than it is to just "spoken language" ?

More general, but more precisely-defined, terms for "spoken interaction", are "talk" and "spoken discourse". Discourse can be defined as "a linguistic structure at the level above that of the sentence" (Harris, 1952), and so "discourse analysis" studies linguistic structures longer than single sentences, and their interaction and

organisation. Features of such structures include anaphors (such as pronouns referring back to something or someone already mentioned) and narrative cohesion and coherence. However, this is not without controversy. The above definition can broadly be described as a "formalist" or "structuralist" approach to discourse (Schiffrin 1994). Others would prefer a more "functionalist" definition, relating to what the discourse is being used for. For example, a single word such as "Stop !" or "Gentlemen" (whether spoken, or on a sign) could be considered as a "discourse" in its own right, even though they have no "structure at a level above a single sentence". An alternative definition of "discourse" which follows this perspective would be "language in use – used to do something and mean something, language produced and interpreted in a real-world context" (Cameron 2001, p13). Of course, there are several other possible definitions of "discourse" as well.

Hymes (1972a,b) proposed a scheme based on the idea of "communicative competence" (combining Chomsky's ideas of linguistic "competence" and "performance") to investigate "rules of speaking". Within Hymes' scheme, there are three levels of "speech unit". At the top level, the "speech situation" is the social context within which the speech of current interest occurs, but includes things other than the speech alone. For example, a school lesson may include writing on the part of various people, gestures, facial expressions and other activities in addition to the actual spoken utterances. At the intermediate level is the "speech event", which is the actual speech (or collection of utterances, dialogue turns, etc.) of interest, within a single speech situation. At the bottom level is the "speech act", which can refer to a single utterance or short set of utterances by a single speaker. "Greeting", "asking a question", "answering a question", "insulting", "apologising" are examples of "speech acts" in the sense that Hymes uses the term. Thus, a "speech event" can be considered as a sequence of speech acts, together with certain additional information. Hierarchies of this nature have been used in some computational models (e.g. Jurafsky et al 1997). Hymes proposed a structure, sometimes called the SPEAKING grid (since the word SPEAKING can be used as a mnemonic for the components of the structure) to describe the content and context of a speech event :

S : setting, where the speech event occurred (time and location)

P : participants who took part in the speech event, and what part each one played in it (e.g. speaker, person being addressed, eavesdropper, ...)

E : ends – the purpose and intended outcome of the speech event.

A : act sequence – the sequence of the speech acts making up the speech event

K : key – the tone or manner of the execution of the event (e.g. serious or joking, sincere or ironic, ...)

I : instrumentalities - the medium of communication (speech, sign language, writing , ...) and language or variety from the participants' repertoires.

N : norms of interaction : the rules for producing and interpreting speech acts within the current framework.

G : genres – the "type" or "class" which the speech act belongs to, and other relevant features of the current act (e.g. is the speaker quoting poetry, or from a standard religious text).

(After Cameron, 2001, p 56)

Schriffrin (1994) and Cameron (2001) have suggested that Hymes' scheme should be used as a "heuristic" for the analysis of speech events, rather than an algorithm for processing them. However, it does provide one possible framework for the systematic description of a conversation in its context.

From a more "sociological" perspective, the important issue is understanding the "orderliness of social interaction" – how the order of such social interactions is produced and how it can be reproduced. This approach to spoken discourse is known as "Conversation Analysis", pioneered by Harvey Sacks. The viewpoint of Conversation Analysis (or C.A.) is that the participants in a conversation are **not** just automata following external rules but create a "social order" actively and continuously through their own behaviour. Thus, in contrast to the theory of speech acts (see below) which might define a given speech act as a "question" because it meets a specified set of criteria, conversation analysis would be concerned about whether the question was followed by an answer (Cameron 2001). However, it does not assume that there are no rules or procedures for the participants to follow – if there were not, the result would be expected to be a chaotic random collection of

utterances. Rather, it is primarily interested in what procedures participants in conversation follow in order to produce normal, "orderly", structured conversations. Some computational approaches make use of the concept of "dialogue moves" or "move-game" theory (Power 1979, Carletta et al 1997) which can be related to these procedures.

One of the key observations of C.A. is the nature of turn taking in dialogue – or, more generally, in "polylogue" conversation. Conversation requires the participants to take turns at speaking. At the risk of stating the obvious, most of the time in any conversation, a single participant is speaking at a time. Although there may be occasions when more than one participant attempts to speak at the same time, such occurrences will normally be treated as a problem – an anomalous, awkward situation or error which needs to be rectified - by all concerned and "speech repairs" will take place – typically, possibly after some apologies and negotiation, one speaker will be allowed to continue whilst the others become silent. Similarly, conversations do not normally include lengthy periods when no participant speaks – in such a case, the conversation will tend to "die". Furthermore, it is not normally the case that a single speaker continues indefinitely – "speaker change recurs". However, neither is it the normally the case that the order of people speaking, or the time for which they speak, is pre-planned. Instead, with the exception of "controlled" situations such as chaired meetings where the chairperson controls who is to speak, the process through which the person speaking changes from time to time is one of continual negotiation between the speakers. Sachs et al (1974), quoted in Cameron (2001), proposed a simple scheme by which turn-taking in a multi-person conversation (or "polylogue") could be governed:

(1) Normally, the current speaker selects the next speaker (e.g. "Would you agree, John ?"),

or, if that does not apply,

(2) The new speaker self-selects (e.g. "If I could just make a comment regarding that ..."), or, if that does not apply,

(3) The current speaker may (but does not have to) continue.

When working with a dataset which has been transcribed and marked-up in an appropriate way, of which the British National Corpus is an example, the concept of speaker turns is a natural idea to incorporate into a model of dialogue. Extensive use of data partitioned into turns is made (although primarily in the context of conversations involving just two participants) in the models used in the remainder of this study.

An approach based in the part of linguistic theory known as pragmatics, which has strong connections to philosophy, is concerned with how language acquires meaning as it is used. There is a distinct contrast here between a purely "formal" or "symbolic" language (such as is used for formal logic, computer programming or for mathematical equations) and "ordinary" natural language which we use for speech. What we say may mean more (or less) than, or something rather different to, the "face value" meaning of the sequence of words used. For example, if one person says to another, "It's cold in here !", the speaker may really mean "You shouldn't have left the door open !". Contextual factors are clearly of great importance to such situations. This field was largely developed by J.L. Austin, John Searle and H.P. Grice and includes the theory of "speech acts" (Austin 1962, Searle 1969) – a topic of relevance to some approaches to modelling dialogue (e.g. Stolke et al 2000). The theory of speech acts and, more generally, how pragmatics relates to discourse, are described in Schiffrin (1994, Chapter 3 pp 49-96 and Chapter 6 pp 190-231) and Cameron (2001, Chapter 6, pp 68-86). Cohen & Perrault (1979) made an attempt to incorporate a speaker's intentions and plans into a computational model of speech acts, which was further developed by Allen & Perrault (1980) (see also Perrault & Allen 1980), eventually leading to the "joint activity" model of Cohen et al (1990).

Even theoretical approaches which can be said to have their primary origins in linguistics have very significant differences.

The "structuralist" approach, following Harris' (1952) definition of "discourse" as "structure within language at levels above the sentence", concentrates on looking for formal regularities and patterns, which can be described as general (or reasonably general) rules. However, there may need to be different rules for "controlled situations" such as a question/answer session in a classroom (e.g. the teacher asks a

question, a pupil responds, then the teacher gives an indication of whether or not the answer was satisfactory, before asking another question, ...) from those relating to situations where the conversation is spontaneous. Labov (1972a) and Labov & Fanshel (1977) examined the "structure of spoken narrative" in detail within the "discourse of therapy" – the type of conversation which arises when a patient is consulting a doctor – in an attempt to discover regularities behind the often higglety-pigglety "surface" appearance of such conversations. E.g. (fictious):

Patient : "I've been suffering from headaches for several days."
Doctor : "Have you also been experiencing indigestion, or other stomach problems ?"

These two utterances may appear rather unrelated, but the doctor may have an insight that problems with digestion often lead to headaches. To quote Labov, "The fundamental problem of discourse analysis is to show how one utterance follows from another in a rational rule-governed way. Within this type of framework, the use of anaphors (e.g. pronouns) and resolution of what they refer to, is an important topic. For example (adapted from Sacks, 1972), in the context, "The baby cried. The mother picked it up.", the use of "it" refers to the same baby and most probably "the mother" is that baby's mother. The resolution of anaphors has also been studied from the perspective of computational syntax (Lappin & Leass 1994, Mitkov, Lappin & Boguraev 2001).

However, many of the more "social" and "contextual" aspects of some of the theoretical approaches to discourse analysis pose much greater problems from the point of view of computational modelling. Although it is clear that such factors do have a major influence over the meaning of utterances in dialogue and how a dialogue or other conversation develops, incorporation of such features into a computational model or system would require a much higher level of linguistic comprehension than is currently possible by a computer. For example, how could a computer be expected to interpret an utterance incorporating irony (e.g. "Well, that was a sensible thing to do !"), or making use of "hidden" knowledge (as in the doctor and patient example above) ? Although their incorporation could potentially lead to a highly sophisticated automatic language understanding system, and to much greater adaptability of such a system to the situation in which it is currently being used, such features are at present

considered not suitable to be included in mainstream computational models of dialogue or used in automatic interfaces.

It may not just be computers or other automated systems which fail to understand or take account of "hidden" or "implicit" knowledge required to correctly interpret the meaning of utterances within a conversation. In many instances, this may also apply to humans ! Clark (1994, 1996) has argued that language use is fundamentally a collaborative activity between participants, and that a crucial issue for the success of a conversation is that the participants understand all the utterances to an extent which is sufficient for the current purposes. A crucial concept for this is one of "grounding" (Clark & Marshall 1981, Clark & Brennan 1991) the conversation - the participants need to be able to agree on a certain minimum amount of shared knowledge and will need to request and give clarifications and explanations until this has been achieved. Of course, such a process requires continuous updating - just because a conversation has been successfully grounded at an early stage does not mean that it will necessarily remain grounded for the remainder of its duration. Subsequent utterances by one or more parties may rely on knowledge not shared by the other participants and so further "grounding negotiations" will be required. Ginzburg (1998, 2001b) has discussed the "unique content assumption" - that individual conversational participants each believe that all the participants have resolved this "grounding knowledge" identically - and to what extent this assumption is valid. Clark & Schraefer (1987, 1989) proposed a "contribution model" for conversations. In this approach, a conversation is composed of contributions, each having two phases : "presentation", where the speaker presents an utterance to the listener, followed by the "acceptance" phase, in which the participants try to verify whether mutual comprehension, i.e. grounding, has been achieved. Cahn & Brennan (1999) noted that, similar to Ginzburg's questioning of the "unique content assumption", even in a "grounded" conversation each participant can only estimate how the other parties understand the context. They went on to apply the model they developed to the human-computer interactions involved in querying a database. Traum (1994) and collaborators (Traum & Allen 1992, Traum & Hinkleman 1992) developed a computational approach to grounding in conversation, using "grounding acts" - special cases of speech acts - and a "grounding grammar" to specify what sort of sequence of such acts will result in a successfully grounded context. Traum &

Dillenbourg (1996) produced a theoretical model of miscommunication - where problems arise from the conversation not being grounded initially - and applied their model to cases of task-oriented dialogues (of the type found in the Map Task corpora). Cohen et al (1990) also modelled task-oriented dialogues - using the concepts of "joint intentions" and "joint commitments" shared between the participants in such dialogues within a "joint activity" model.

Although, as can be seen from the above descriptions, the distinct theoretical approaches to dialogue and discourse may overlap to some extent, in many cases they differ in the "scale" on which they focus as well as in their philosophy of approach. Pragmatics works on a large scale – the context of the conversation – whereas the more "structural" approaches tend to focus-in on the rules governing the relationships between successive utterance or dialogue turns at a much more local level. Computational approaches to modelling dialogue can also work at different levels. Alexandersson et al (Alexandersson & Reithinger 1997, Alexandersson et al 1998) identify four different levels within negotiation dialogues : (1) the whole dialogue; (2) dialogue phases, with each dialogue potentially having a greeting phase, a negotiation phase and a closing phase; (3) individual turns within a single phase; (4) "dialogue acts", such as requests, statements, acknowledgements, back-channels, within each turn. This has been applied within the context of the Verbmobil project (see section 3.3 below).

As noted in chapter 2, some approaches from psycholinguistics include concepts which can readily be incorporated into a computational or statistical model. For example, Walker (1996, 1998) proposed the use of a cache model (introduced in section 2.2.2 and further discussed in chapter 5) of approximately the most recent 3 sentences, to explain certain psycholinguistic phenomena relating to humans having limited attention whilst participating in discourse This is proposed as an alternative to the "stack model" (again, by analogy with a "stack" in computer architecture) of Grosz & Sidner (1986, 1998) and Grosz et al (1995). The cache model assumes that the "attentional window" of a participant in a dialogue is largely restricted to the words and utterances most recently spoken (an assumption backed-up by several experimental studies, see e.g. Fletcher 1994, Baddeley 1986), whereas the stack model assumes that the attentional memory is arranged hierarchically, in (say) a tree

structure, so that items and entities dealt with in asides, etc. are erased from the attentional memory once the aside is complete, and attention returns, without difficulty, to the items and entities previously under discussion. The issue "without difficulty" is a key point disputed by Walker (1996, 1998). The prevalence of redundancy in dialogue, where information which had apparently already been agreed between the participants is repeated or "refreshed" (Walker 1992, 1993), is central to her arguments in favour of the cache model. The prevalence of "reprise utterances" (Ginzburg & Sag, 2001) - where the last utterance of the previous speaker is repeated (possibly with different intonation or emphasis) by the current speaker - and clarification requests (Purver, Ginzburg & Healey, 2002) in dialogue, plus the fact that such requests are normally clarified very promptly, are features which provide evidence that a participant holds only the most recent part of the history of the conversation in reliable short-term memory (i.e. a cache). Furthermore, such clarification requests (and the subsequent responses) provide one means for the participants to "ground" the conversation (Purver, Ginzburg & Healey, 2002). Ginzburg (1998) also discussed the use of "clarification utterances" (whether specifically requested or not) in dialogue from the perspective of formal semantics. Cohen et al (1990) found that amongst the natural outcomes of their "joint activity" model for task-oriented dialogues were the kind of "discourse intentions" (Grosz & Sidner 1986, Litman & Allen 1990) which underlie the "discourse markers" - signals such as backchannels, clarifications, elaborations and confirmations - which are so common in dialogue (Oviatt & Cohen 1991).

Similarly, the psycholinguistic concept of "priming" (Bodner & Masson, 2003, Meyer & Schvaneveldt, 1971) - where recent occurrence of a word related in meaning or sound to the current "target" word improves a listener's recognition accuracy or speed of response for that target - is analogous to the trigger pair models discussed in section 2.2.1.

3.3 Previous Approaches to the Modelling of Dialogue in Speech Technology

Most early research on the statistical modelling of dialogue tended to relate to the relatively controlled situations of task-oriented dialogue, such as the Map Task (Anderson et al 1991, Taylor et al 1998, Wright 1998), the content of the VERBMOBIL corpus (Mast et al 1996, Warnke et al 1997, Reithinger et al 1996, Reithinger & Klesen 1997, Samuel et al 1998) and of the ATR Conference corpus (Nagata 1992, Nagata & Morimoto 1993, 1994, Kita et al 1996), which is a subset of the ATR Dialogue Database (Ehara, Ogura & Morimoto 1990).

As noted in section 1.5, the Map Task is one where one participant is trying to direct the other from some specified starting point (A) to a specified destination (B). Each participant has a map, but these two maps are not identical. The starting and end points are common to both maps, as are some other landmarks. However, each map will also contain some landmarks not marked on the other. The task requires negotiation between the two participants so that the "director" eventually describes a route from A to B which is comprehensible to the "directee". The Map Task corpora (Anderson et al 1991, Carletta et al 1997) are therefore clearly examples of collections of task-oriented dialogues. The Canadian English version of this – the DCIEM Map Task corpus (Bard et al 1995) has been modelled (Taylor et al 1998, Wright et al 1999) using a scheme employing Power's "move-game" theory (Power 1979, Carletta et al 1997). The entire conversation is divided into "games", each of which has a specific goal. Individual "dialogue moves" are classified as one of 12 categories, and of these 6 types are also used to specify distinct classes of games, according to the first significant move within that game. These categories are "instruct", "check", "yes/no query", "wh- query", "explain" and "align". The other types of move are "acknowledge", "clarify", "reply yes", "reply no", "reply w (other reply)"and "ready" (used to indicate that the previous game has been completed and that the current participant is ready to commence the next game). They developed a "4-gram" dialogue model (Taylor et al 1998) which made use of the role of the current speaker ("director" or "directee"), the type of the dialogue move last made by the other speaker and the role ("director" or "directee") of the speaker who made the last dialogue move (which could be the current speaker) to predict the dialogue move to

be made next. Just as several sources of information (e.g. predictions from an acoustic model and from a language model) may be combined in an standard automatic speech recognition situation, the predictions of their dialogue model were combined with information from an intonation model and from an "ordinary" speech recogniser in order to obtain a "best" prediction of the next dialogue move. A separate language model was used for each type of dialogue act, so that (for example) yes/no questions were modelled separately from acknowledgements (Taylor et al 1998). Their work also made use of prosodic information in the speech signal to constrain the models being applied to each utterance (Taylor et al 1996, 1998, King 1998, Wright et al 1999, Wright 2000).

The VERBMOBIL corpus (Wahlster 1993, 2000) is based on over 1000 spoken dialogues, of which approximately 300 were manually tagged for the appropriate dialogue acts. The aim of this project was to facilitate communication between two human users, who are native speakers of different languages, by translating parts of their speech on request. The corpus used in the later stages of the project (Alexandersson et al 1998, 2000) was comprised of dialogues relating to the negotiation of the date for a business meeting and making plans for travel, accommodation and entertainment. Their approach to modelling the dialogues is based on a hierarchy of dialogue act types, with a decision tree structure used to determine which type and sub-type any given dialogue turn should belong to. The system used 18 primary types, with a total of 42 sub-types, of dialogue act (Jekat et al 1995, Alexandersson et al 1998). Multi-layer perceptron neural networks were trained to recognise the boundaries between dialogue acts in utterances and N-gram models of sequences of dialogue acts, trained on the manually-tagged dialogues, used to predict the next dialogue act (Reithinger 1994, Alexandersson et al 1995, Reithinger et al 1996, Warnke et al 1997). In parts of their studies, a stochastic context-free grammar was used as a means of recognising the plan within a dialogue (Alexandersson & Reithinger 1997). Their work showed considerable success in predicting the next dialogue act in a sequence. One of their aims was to use this to reduce the search space of the word recogniser (Alexandersson et al 1995), although the extent to which this was successful does not seem to have been made clear in the published reports.

In the work on the ATR Corpus (Nagata 1992, Nagata & Morimoto 1993, 1994, Kita et al 1996), simulated dialogues between a secretary and a questioner at an international conference were modelled using an ergodic Hidden Markov Model. This included a model based on "trigrams of speech acts" (rather than trigrams of words).

Stolke et al (Stolke et al 2000, Jurafsky et al 1997, 1998) have modelled "dialog acts" – such as statements, questions, backchannels, agreements, disagreements and apologies ; roughly equivalent to the "speech acts" defined by Searle (1969), the "conversational move" defined by Power (1979) or the "adjacency pair part" used by Schegloff (1968) and Sacks, Schegloff & Jefferson (1974) – of 1155 conversations, (approximately 1.4 million words in 200000 utterances) within the SWITCHBOARD corpus (Godfrey, Holliman & McDaniel 1992) statistically, regarding the "discourse structure" of the conversation as a Hidden Markov Model (Rabiner & Juang 1986) and each "dialog act" as "observations" resulting from the states of the model. Their model uses lexical, collocational and prosodic cues to detect and predict dialog acts. They used a database of conversations, hand-labelled for dialog acts, to train and evaluate the models, in order to combine speech recognition and dialogue modelling probabilistically with the aim of improving the accuracy rate of both. Their system performed quite well at automatically labelling dialog acts, achieving 71% accuracy when working with word transcripts of the conversations, compared with human performance of 84% accuracy and a chance baseline of 35%. When combined with an automatic speech recognition system, it gave a dialog act recognition rate of 65% accuracy, and a small reduction in word error rate over the baseline of the speech recognition system alone. When using an "Oracle" for selecting the dialog act, they obtained a 13.0% reduction in perplexity and a 2.2% reduction in word error rate over the baseline. Their model allowed for 42 distinct categories of dialog act considered appropriate for the content of the Switchboard corpus. They also attempted to use a cache model of dialog acts and a maximum entropy model based on constraints on dialog act sequences, but the results they obtained were disappointing (Stolcke et al 2000).

The TRINDI (Task oRiented INstructional DIalogue) Project (Larsson & Traum 2000) modelled route planning dialogues, considering both human-human and human-machine dialogues. The computational model used was based on the concepts

of "information state" (the information needed to distinguish a given dialogue from others - sometimes called "mental state" - at a particular time) and "dialogue moves" which update the current information state. The project produced a toolkit for developing the "dialogue manager" component of an automatic spoken dialogue system.

Young (2000, 2002) has modelled a spoken dialogue system as a Partially Observable Markov Decision Process (POMDP), so that the model has minimal dependence on explicit rules being programmed or "hard-wired" into it, but can readily learn from observation of data via a reinforcement learning strategy. Each dialogue act required incurs a small penalty (or "negative reward") - on the basis that, in a spoken dialogue system, longer dialogues are likely to cause more frustration to the human user and higher costs to the provider of the system. However, successful completion of the dialogue, to the user's satisfaction, results in a large positive "reward". The sequence of interactions between the user and the system from initialising the dialogue to its termination is known as a "dialogue transaction". The learning strategy is such that, based on experience from training examples, the system modifies a "policy matrix" giving probabilities that the current input is u given that the system is currently in state i – and a "transition function" – giving probabilities of going to any new state j given that the system is currently in state i and that the current input to the system from the user is u – in order to maximise the expected total net reward for the complete dialogue transaction. (However, unless we are at the end of the dialogue transaction, we cannot know the total net reward, only estimate it.) Approaches to performing the required optimisation, either by dynamic programming or by sampling methods such as Monte Carlo techniques, are discussed in Young (2000), who notes that the dynamic programming approach has the drawback that it requires complete knowledge of the systems' transition function beforehand, whilst using a sampling approach to learn from examples on-line as they occur has the disadvantage that many such training cases will be necessary to obtain a near-optimal policy. With this in mind, Scheffler and Young (1999, 2000, 2002) employed a model to simulate users' behaviour. Walker and Young (2003) have followed a different approach – using "Wizard of Oz" examples (where the user is led to believe that he/she is communicating with a computer when, in reality, it is another person) with reinforcement learning to train a dialogue management system.

Recently, He & Young (2003) have developed a generalization of Hidden Markov Models (called "Hidden Vector State Models") which accommodates an efficient representation of hierarchical structure in such a way that long-range dependencies can be learnt from unannotated training data. This in turn allows semantic ambiguities to be resolved in parsing. The results of their experiments using the ATIS-3 1993 and 1994 Air Travel databases have been encouraging, with their system giving fewer errors in "goal detection" tasks and better accuracy in source and destination "slots" than a general finite state tagger.

Halliday & Hasan (1976) and Clark & Haviland (1977) proposed that sentences or utterances in dialogue have an "informational structure", separate from the syntactic structure. Part of the utterance deals with "given" information – information which is already shared or agreed between the participants – and "new" information, being imparted for the first time by the current speaker to the current listener. "Given" information tends to occur near the start of the utterance, whereas "new" information tends to occur later. A preliminary attempt to exploit this within dialogue modelling was made by Mateer & Iyer (1996). They applied a rule of thumb – that the part of a sentence before the main verb is "given" information, whilst the part after the verb is "new" - to data from the Switchboard corpus, finding that the vocabulary and lexical distribution were significantly different between the two parts. However, this could, in part, be due to the syntactic structure of English – particularly in cleft sentences - rather than necessarily being due to the informational structure. Nevertheless, the distribution of certain types of words would be expected to be different between the two parts of an utterance. Pronouns (and other anaphora) requiring a prior reference would be expected to mainly occur in the "given" portion of the utterance, whereas verb forms relating to the semantic content will tend to be found in the "new" section. Broadly speaking, this was what was found by Mateer & Iyer (1996). They went on to develop, train and test separate bigram language models on the full sentences, on the parts of the sentences which came before the main verb and on those parts of the sentences which came after the main verb, finding major differences in perplexity according to which material each model had been trained and tested on. The ideas of "given" and "new" (or "topic" and "focus") information within dialogue utterances have also used within a statistical framework, using a "self-organising map" type

neural network (Kohonen 1982, 2001), by Lagus & Kuusisto (2002), applied to a corpus of Finnish dialogues.

This work has been generalized in the Interact project for Finnish dialogues, where Jokinen et al (2002) have employed a hybrid approach using an architecture based on "agents", incorporating a self-organising map neural network for dialogue topic recognition (Lagus & Kuusisto 2002) and a "constructive dialogue model" (Jokinen et al 2001) for modelling sequences of dialogue acts. Targeting a language model to the current topic is particularly important for a strongly inflected language with a highly flexible word order, of which Finnish is an example. For such languages, short-range models such as trigrams will be of much less value than for languages with a much stricter word order, such as English. However, it might be expected that an approach making use of topic identification might also be of value, if less crucial, if applied to strict word-order languages like English. Kurimo & Lagus (2002) carried out a comparative study, applying this type of approach to data in both Finnish and English, finding an improvement in perplexity in the models over a trigram baseline for both languages. However, the datasets were news-text in Finnish and patent abstracts in English, so it is not possible to judge the benefit of such an approach for English dialogue from that study.

Of course, not all computational approaches to modelling dialogue are statistically-based. For example, the ROSSInI – Role Of Surface Structural Information In dialogue – project (Ginzburg 2001a, Purver et al 2001, Purver 2002) takes a perspective based in semantics, using the framework of Head-Driven Phrase Structure Grammar (HPSG) (Ginzburg & Sag 2001). This HPSG approach has been applied to several aspects of dialogue – in some cases using dialogue data from the BNC. It has been applied to resolving fragments in dialogue (Ginsburg et al 2001) and the resolution of ellipsis and anaphora (Lappin & Gregory 1997, Ginzburg 1999). Subsequently, the work has been extended to deal with incomplete utterances which may not contain verbs – so called "non-sentential utterances" (NSUs) – in the projects "Phrasal Utterance Resolution in Dialogue" (Gregory 2001) and PROFILE : Processing and Resolution Of Fragments In dialogue (Ginzburg 2003, Fernández & Ginzburg 2002). Such fragments are particularly common in dialogue, accounting for possibly as much as 11% of utterances in a sample from the dialogue material from

the BNC (Ginzburg 2003, Fernández & Ginzburg 2002) or even 30% of utterances between members of a single family (Van de Waijer 2001, quoted by Ginzburg 2003). Ebert et al (2001) used a "template-filler" approach within an HPSG framework, recycling syntactic and phonological information from parsing and interpretation, as an efficient means of generating full paraphrases for fragmentary utterances within dialogue. Ginzburg et al (2001b) have attempted to use HPSG to incorporate "conversational move types" into a grammatical analysis of conversation, whilst Purver et al (2001, 2002) have investigated clarification requests – another very common feature of human conversation - in dialogue within an HPSG analysis framework. This has been extended to the processing of unknown words by a dialogue system, via requests for clarification, by Purver (2002). The problem with most techniques based on purely syntactic or semantic approaches to Natural Language Processing (NLP) is that they do not normally give quantitative predictions of word probabilities which might be used to assist word recognition or choose between multiple interpretations. Nevertheless, some connections between the NLP and statistical approaches may come out of an analysis of the results of statistical methods applied to dialogue, and consideration of how and why these methods work as they do.

3.4 Aims and Hypotheses for the Remainder of this Study

In this chapter, we have discussed ways in which dialogue is believed to be different to written text and to spoken monologue. Yet, as previously noted, most language models for automatic speech recognition or understanding systems are trained on text material, broadcast news transcripts and similar materials - and for that matter, most acoustic models are trained on read speech from similar types of sources to these. However, if normal conversational speech – and dialogue in particular – really is quite different from such source data, it would seem far more appropriate that language and acoustic models be trained on the same sort of material to which it is intended that they will be applied. This issue has been discussed by Rosenfeld (2000b).

In the remainder of this study, the aim is to study, from the point of view of statistical language modelling, just how dialogue material – or that least the sample of modern British English dialogue contained within the BNC – differs from text. How does the vocabulary variation and distribution differ between similar amounts of text and dialogue ? How well do various techniques regularly used in statistical language modelling – in particular, the types of model described in chapter 2 – work for dialogue material in comparison to text material ? Can we give any linguistic interpretation of some of the results from statistical language modelling experiments on dialogue and text material respectively ? Can any of these experiments shed any light on structural aspects (as opposed to purely lexical and/or topic based properties) of dialogue. Such questions will be addressed in the next four chapters of this thesis.

Chapter 4 Dialogue Material in the British National Corpus

4.1 The British National Corpus (BNC)

The British National Corpus (Burnard 1995, BNC 2001), henceforth referred to as the BNC, is a large database of modern British English, compiled between 1991 and 1994. The project was a collaboration between several academic institutions, commercial publishers and the British Library, funded by the UK Department of Trade & Industry (DTI), Science & Engineering Research Council (SERC), the British Library and the British Academy.

The BNC is composed of both written material (totalling around 90 million words), transcriptions of spoken monologue (around 1.9 million words) and transcriptions of spoken dialogue (around 7.7 million words). It was designed with a view to represent as wide a range as possible of modern British English. With this in mind, the written component includes a diverse variety of text, including letters (both published and unpublished), essays by both school and university students and extracts from both national and regional newspapers and specialist periodicals, in addition to material from novels and academic books. Similarly, the spoken material was collected from various different sources. Some were "context governed" situations, such as business meetings, medical consultations, school lessons and interviews, whilst others were taken from recordings of spontaneous informal conversations. This latter material was recorded by volunteers who were selected in a way intended to be representative of the population of modern Britain in terms of age, region and social class in a demographically balanced way and is therefore referred to as "demographically sampled" material.

4.2 Spoken Material (including Dialogue) within the BNC

The spoken component of the BNC (about 10% of the whole corpus) consists of data from 863 sources, of which 153 (approximately 4.2 million words) are "demographic material" (spontaneous everyday speech) and the remaining 762 (approximately 6.15 million words) from "context governed sources" (such as business meetings). The audio recordings have been deposited at the National Sound Archives of the British Library.

Of the 863 spoken data files, 672 are listed as being "dialogue" (although many of these contain conversations where there are more than two speakers, and many have a single speaker doing the majority of the talking). These "dialogue" files occupy 144666 kBytes (transcribed, in SGML marked-up format) and contain a total of 7760753 words in 888535 sentences.

To quote Burnard (1995), "The importance of conversational dialogue to linguistic study is unquestionable : it is the dominant component of general language both in terms of language reception and language production."

The "demographic" part of the spoken corpus was collected using a "demographic sampling" of the population of British English speakers in the UK – designed to contain data from 124 speakers representative of the UK population in terms of age, gender, social group and region. The selected individuals recorded all their conversations over a period of two or three days, recording details of these conversations in a notebook. Other participants in these conversations gave their permission for the data to be used prior to its inclusion in the corpus.

The "context-governed" part of the spoken corpus recorded speech from four broad categories of sources in approximately equal amounts : (i) educational and informative situations, such as news broadcasts, classroom discussions and tutorials; (ii) business events, such as sales demonstrations, meetings, consultations and interviews; (iii) institutional and public events such as council meetings and parliamentary proceedings; (iv) leisure events, such as sports commentaries, club meetings and radio phone-ins.

As part of the original BNC project, the original speech recordings were orthographically transcribed, word-tagged (for parts of speech, etc.) using an automatic system and "marked-up" using SGML (Standardised General Mark-up Language). The SGML mark-up not only indicates which speakers are involved in a conversation, but also their social relationship and the place and context in which the conversation occurred. The mark-up also indicates disfluencies and occasions where the speech of the two current participants overlapped. However, in the studies described in this thesis, the transcriptions were pre-processed, removing the SGML tags and only looking at sections of transcription involving a single pair of speakers. We have chosen to define a "dialogue" strictly as comprising a sequence of consecutive utterances from exactly two speakers, so a new logical dialogue commences whenever a new speaker joins (or a previous speaker, not one of the current pair, rejoins) the conversation. In this study, all the portions of overlapping speech have been made to appear sequential, all non-transcribed disfluencies have been removed and any turns which became empty as a consequence of these changes deleted. The word-tag information has not been used.

4.3 Descriptive Statistics

As noted above, many of the "dialogue" files in the BNC contain contributions from more than two speakers and thus a single file may contain several logical "dialogues" (as defined in section 4.2 above). In fact, the length of the files varied extensively – ranging from very short exchanges between two speakers to long debates involving many speakers – in a highly skewed manner. Some summary descriptive statistics relating to the content of the BNC dialogue data – in terms of files, dialogues, turns and words – are shown in Table 4.1 below.

Despite the modal number of "dialogues" (according to the above definition) per file being 1, the median number was 20.5 and the 70^{th} and 80^{th} percentiles were 66.1 and 154.4 respectively. One file contained over 3100 distinct dialogues ! Such unusually high values are, in part, due to the way in which a "dialogue" has been defined– which was largely for simplicity and computational convenience - in this study. A

consequence of this definition is that a conversation involving (say) just 3 speakers could end up consisting of a very large number of individual dialogues, according to this definition. Each time there is a change to the "current pair of speakers" (e.g. speaker B is replaced by speaker C, whilst speaker A continues), a new dialogue results. In all, the 672 BNC "dialogue" files contained a total of 91650 dialogues according to our definition.

	Minimum	Mode	Median	Mean	Maximum
Dialogues in a file	1	1	20.5	136.4	3115
Turns in a dialogue	1*	2	2	6.19	2326
Words in a dialogue	1*	9	19	79.36	21123
Words in a turn	1	1	5.5	13.0	18575
Proportion of words in a dialogue by first speaker	0.00*	0.500	0.500	0.499	1.00*

Table 4.1 : Some summary descriptive statistics for the BNC dialogue material.
* These "dialogues" are clearly not true dialogues. Such "pseudo-dialogues" are probably due to an error in the transcription or mark-up within the BNC material, or to the deletion of turns which have become "empty" during the removal of disfluences in pre-processing.

Similarly, although less directly affected by the way in which we defined a dialogue, the number of turns within a single dialogue, the number of words in a dialogue and the number of words in a turn all showed wide variation in a highly non-normal, skewed manner. The modal number of turns in a dialogue was just 2, as was the median. The mean number was 6.2 turns per dialogue and the 90^{th} percentile 8. However, one dialogue contained as many as 2326 turns. The number of words in individual dialogues showed even greater extremes. The modal number was just 9 and the median 19, but the mean was 79.4 (larger than the 80^{th} percentile, which was 55) and one dialogue contained 21123 words ! Thus, although most dialogues in this corpus are quite short, the distributions of words and turns have very long "tails".

The dialogues also varied considerably in their "balance" - the proportion of the words spoken by each of the participants. However, on the whole, there was no evidence that either the first or second speaker to enter that dialogue tended to dominate it. The mean, mode and median proportion of words spoken by the first speaker were all 0.50. 80% of the dialogues had the first speaker contributing between

11.7% and 88.9% of the total words of that dialogue and such cases were deemed to be "reasonably well-balanced".

4.4 Lexical Distribution

The dialogue material within the BNC was found to contain 49 989 distinct word types (i.e. words with distinct orthographies). This contrasts with the figure for an equivalently-sized sample (of 7 million words) of data from the written text material in the BNC, which was found to contain 104 827 distinct word types. A larger sample (80 million words) of written text contained a total of 352 860 distinct types. Thus, it can be seen that the dialogue material uses a considerably smaller vocabulary than the written portion of the BNC. As might be expected, there is also a marked difference in the increase in coverage of material with increasing lexical size (i.e. investigation of what proportion of the material is accounted for by use of a lexicon of the N most common distinct words) between the written text and dialogue datasets. This is illustrated in Figure 4.1 below. For extremely small lexica (less than 9 distinct words), a slightly higher proportion of text material is covered than dialogue material – this may be due to the prevalence of very common articles ("a", "the") and conjunctions ("and") in text, which are less common in dialogue – in the latter case partly due to sentences typically being shorter in dialogue than in text. However, at larger lexical sizes, a higher proportion of the dialogue material is covered by the appropriate lexicon of any specified size. This finding is not so surprising, since it would generally be expected that simpler words are particularly common in dialogue whereas relatively rare, esoteric words are more likely to appear in written material than in speech.

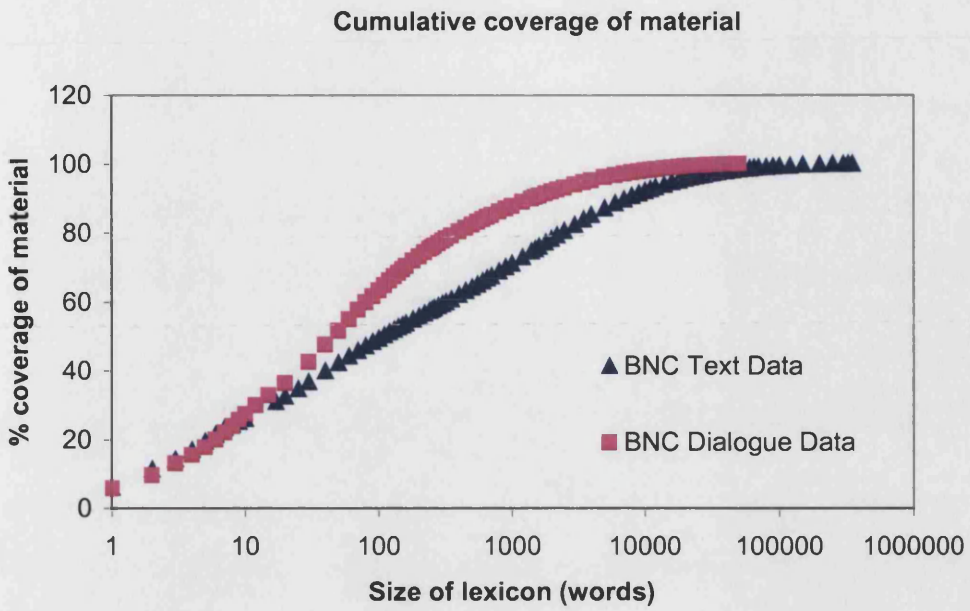


Figure 4.1 : Graph contrasting coverage of material by lexica of various sizes for BNC text & dialogue material

The smaller lexicon (fewer than 50 000 distinct words) relevant to the dialogue material is small enough to be considered as a "closed system" and hence provides two advantages with regard to performing statistical language modelling : firstly, there are no significant issues about how to deal with out-of-vocabulary words, and secondly there should be fewer problems in smoothing n-gram language models since there are fewer words which occur very few times (particularly in the case of words which occur only once in the available data – so-called "hapax legomena").

4.5 Dialogue Reduced-Turns (DRT) Dataset

Some of our preliminary investigations (as described in section 4.3 above) indicated that, although identified in the file descriptions as "dialogue material", a considerable portion of the spoken component of the BNC consisted of very long speaker turns – in effect, to a close approximation, monologue, possibly interspersed with occasional comments or questions. Such material is not what we are primarily interested in studying – this project aims to investigate what is particularly distinctive about highly interactive dialogue. It might be expected that long dialogue turns have some properties which are not so different from ordinary text material – in particular, a long turn gives opportunity for any single topic to become firmly established. In such

cases, it would be expected that topic-related content words would be quite significant – both as a fraction of the total words in a turn or dialogue, and in terms of relations between words of the type exploited by cache a trigger models. Whilst this is interesting in its own right, it is also important to study word dependencies which are consequences of the dialogue structure rather than of the content or topic. Thus, it was decided that, in addition to studying the properties of, and word dependencies within, the "ordinary" (i.e. unfiltered) dialogue material in the BNC, to focus specifically on pairs of relatively short successive turns from within the same dialogue. Such pairs were believed to be more typical of the type of turns likely to occur in highly interactive dialogues, of the kind appropriate to both human-human and human-machine interactions. Furthermore, these short pairs of turns are expected to have relations between words which are much less dependent on the topic and content than those within very long turns.

The set of files in the BNC listed as "dialogue material" were removed of their SGML mark-up and divided into pairs of successive speaker turns – effectively restricting each "dialogue" or "document" to consist of just two turns. Only pairs totalling fewer than 200 words were retained, yielding a set of approximately 470000 pairs. The resulting data was called the Dialogue Reduced Turns (DRT) dataset, to distinguish it from the full set of ordinary "dialogue" material in the BNC.

4.6 Simple Statistical Language Models

4.6.1 Trigram Models for Ordinary Dialogue Data

Simple trigram language models were constructed for the dialogue data from the BNC using the CMU-Cambridge Language Modelling Toolkit (Clarkson & Rosenfeld 1997). The complete dialogue dataset was divided into 10 sets each of training data (approximately 7 million words each), evaluation data (approximately 380 000 words each) and test data (approximately 380 000 words each) for use in a 10-fold cross-validation procedure. Cross-validation is a method which attempts to make the best possible use of the available data for both training and testing whilst keeping the training and test material used in any one training-test pair ("rotation") totally separate. For each experiment, a 10-fold (i.e. using 10 rotations) cross-validation

procedure was employed such that, in each rotation, 10% of the available data (approximately 770 000 words) was held back for evaluation and testing, with the remainder being used for training the model(s). A different 10% portion was used for evaluation & testing for each of the rotations.

For comparison, trigram language models were also constructed from samples of randomly-selected text data from the BNC of sizes 5, 10, 20, 40 and 80 million words respectively. The 50 000 word lexicon covering the dialogue material was used throughout. Good-Turing smoothing (Katz, 1987) was applied in each case in order to allocate small non-zero probabilities to possible trigrams which were not found in the training data, with singleton cut-off being used for computational convenience.

Across the ten cross-validation rotations, the mean perplexity for the trigram models trained on the 7 million words of dialogue data was found to be 186 (with a variation of $\pm$ 12 across the ten rotations). The trigram models trained on text data showed significantly higher perplexities (see figure 4.2). For example, a model trained on 10 million words of text gave a perplexity of 389 (with a variation of $\pm$ 17 across the rotations). (Although this perplexity value may appear high in comparison to published values for other corpora, it is consistent with the values obtained by Clarkson & Robinson (1998) and Clarkson (1999) for a trigram model trained on 105 million words – as quoted by Clarkson (1999) – from the BNC, without distinction between text and spoken material.) As expected, the mean perplexity of a text-trained model was lower the larger the set used for training was. The variation of perplexity (with respect to held-back data) with the size of the training set is shown in figure 4.2. If the trend shown continues, it would appear that over 500 million words of plain text training material would be needed to approach the perplexity obtained from using just 7 million words of dialogue training material. Results of this type, where the perplexity of a model is particularly sensitive to both the nature of the material on which it is trained and on which it is tested, have been noted by Rosenfeld (1996, 2000b). For modelling casual telephone conversations, Rosenfeld (2000b) stated that using 2 million words of transcripts from appropriate telephone calls as training material would be better than use of 140 million words from transcripts of TV or radio news broadcasts. Even changing the material within what at first sight might appear to be the same domain can make a big difference : an experiment showed that a language

model trained on material from text from the Dow-Jones newswire had its perplexity almost doubled if tested on material from the Associated Press newswire, compared with Dow-Jones text from the same time (Rosenfeld 1996), despite these two sources generally being considered to be very similar ! Both the text and spoken portions of the BNC contain material from a very wide range of sources and on a large variety of topics. This at least partly explains why the perplexity figures for text language models trained on BNC data are so high. Clarkson and Robinson (1998) found that a trigram model trained on about 100 million words of BNC data (mixed text and spoken) had a perplexity of 277.5, compared with the corresponding figure of 134.4 for a model trained on 130 million words from the much more homogeneous Broadcast News Corpus.

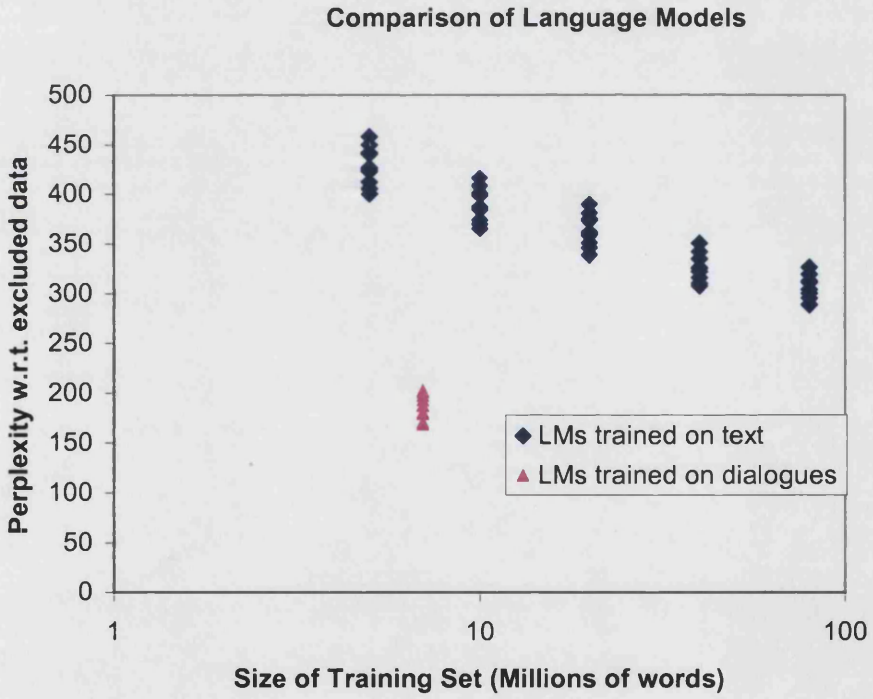


Figure 4.2 Variation of perplexity of trigram language models (with respect to excluded data) with size of training corpus for text data, with corresponding values for models trained on dialogue. The multiple points for models trained on the same size of training corpus represent results from the 10 distinct cross-validation rotations for training corpora of that size.

The variation of trigram model perplexity with the size of the corpus used to train the model for BNC text data prompted the creation of the TEQ (or "Text Equivalent")

dataset – a subset of the BNC text material of equivalent size (approximately 7 million words) to the BNC dialogue dataset used to train the dialogue language models. Using this TEQ dataset, rather then the whole written text portion of the BNC, to train language models for text enables a more direct comparison to be made between models trained on dialogue and those trained on text – differences (in model perplexity, etc) between similar models trained and tested on the two different types of material should then primarily be due to the nature of the material rather than the size of the datasets used for training.

4.6.2 Trigram Models for DRT Data

In a similar manner to that used for modelling the ordinary dialogue data (see section 4.6.1 above), trigram language models were constructed for the contents of the DRT dataset using the Cambridge-CMU Language Modelling Toolkit. Once again, 10 sets each of training (approximately 423000 turn pairs in each set), development (approximately 23500 turn pairs) and evaluation (approximately 23500 turn pairs) were created for use in a 10-fold cross validation procedure. In every case, the dialogue lexicon of approximately 50000 words was used, and Good-Turning smoothing with singleton cut-off applied.

For purposes of comparison, three separate sets of trigram models were constructed : one only for first turns of pairs, one only for second turns of pairs, and one for both turns together, and the perplexity of each model with respect to held-back data computed. Averaged across ten cross-validation rotations, the results for the models constructed from single turns were almost identical : 187.61 for the first turns of pairs and 187.69 for the second turns of pairs. These contrasted with the substantially higher figure of 289.61 obtained for the model trained on both turns of the pairs – treating the dialogue material very much as though it were ordinary text. The variability of perplexity across the 10 cross-validation rotations was also rather higher for the case where both turns of each pair were considered together (maximum 477.15, minimum 150.76) compared with the models for first turns only (maximum 265.62, minimum 145.81) or for second turns only (maximum 261.75, minimum 145.58). These very variable, often higher, figures are probably due to probabilities

given by the model being distorted because of replication of turns between successive pairs.

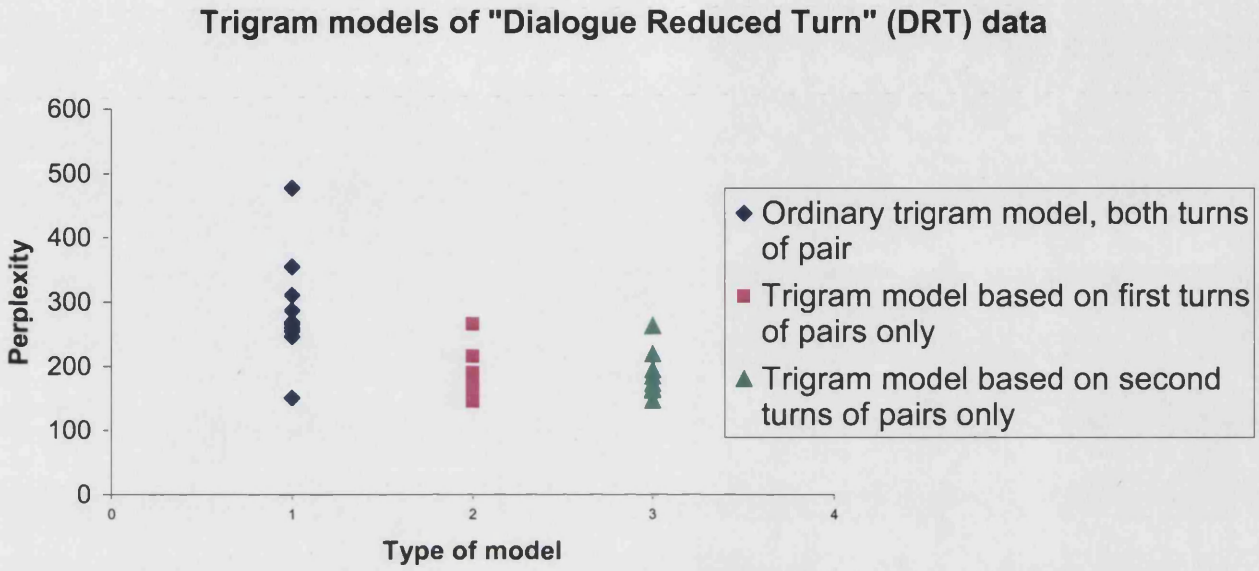


Figure 4.3 Comparison of trigram models for DRT data, modelling first turns of pairs only, second turns of pairs only, and both turns of pairs together respectively.

The concept of a "turn pair" is not meaningful for ordinary text data, so no comparison could really be made between perplexities of DRT data and of ordinary text data. However, the perplexities of the trigram models for single turns of pairs are comparable to those for ordinary dialogue data (see section 4.6.1 above).

For an automated spoken dialogue system, the machine will be required to predict the content of the second turn of a pair (i.e. the user's turn) based on knowledge of the first turn of that pair (i.e. the machine's own turn). Hence, modelling the second turns of pairs is of particular interest and so the trigram model for second turns of pairs was used both as the baseline for comparison and for purposes of interpolation with other models in later experiments (see subsequent chapters).

Chapter 5 Experiments Using Cache-Based Language Models

5.1 Overview

As discussed in section 2.2.2, the results of previous modelling experiments with text data have shown that cache models are a simple but effective means of tracking how lexical likelihoods of words vary with the topic of the current document (Kuhn & De Mori 1990, Iyer & Ostendorf 1999), thus allowing a form of adaptation within a language model. A typical cache model maintains a history of recent words used in the current document, and estimates a dynamic unigram language model using solely those words. This is then interpolated with a static trigram model built from a much more topic-independent text. Experimenting with various cache sizes, Clarkson & Robinson (1997) found that a cache of 500 words performed better than any other.

It might be expected that statistical modelling of the BNC dialogue material would also benefit from a similar approach: that the likelihood of words used in a dialogue would be affected by which words had been used earlier in the same dialogue. In particular, it would seem likely that the content of the second of a pair of consecutive dialogue turns would be closely related to that of the first turn of the same pair.

To evaluate the utility of cache models in the context of dialogue, we constructed a set of cache models. In one set of experiments, we used either a fixed size-cache ("F"-type experiments) of 500 words, or a cache consisting of the current and the previous dialogue turn ("T"-type experiments), or a cache consisting of the current and the previous sentence ("S"-type experiments) applied to dialogue data from the BNC, otherwise treating the dialogue material as though it were ordinary text. In each case, the cache was "flushed" (reset) at the start of each new dialogue – i.e. each dialogue was treated as a completely distinct source of data. Each cache model was based on unigram statistics (i.e. word counts) within the cache. For purposes of comparison, we also did parallel experiments for cases "F" and "S" on an equivalent-sized sample of ordinary text material from the BNC. The effect of varying the cache size for "F" type experiments was also investigated.

In another set of experiments, a cache model applied to pairs of dialogue turns (from the DRT data set), with a sliding window of at most 500 words was used to construct two caches : one for the first turn of the pair, the other for that portion of the current (second) turn which had already been encountered. The caches were reset at the start of each document (namely each new pair of turns). The resulting cache models were interpolated with a trigram model trained on the content of second turns of pairs. Optimal interpolation parameters were learned using the *interp* program in the CMU Language Modelling Toolkit, which performs Expectation-Maximisation (EM) training (Dempster, Laird & Rubin 1977, Jelinek 1990) on matched probability streams. The interpolation parameters were trained on the Development test sets of the dialogue data and applied to the Evaluation sets using a 10-fold cross-validation procedure.

5.2 Cache Experiments on Ordinary Dialogue Data from the BNC

5.2.1 Comparison of Fixed Size, Turn-Based and Sentence-Based Caches

As described above, in a first attempt to investigate the utility of cache-based models in modelling the dialogue material within the BNC, the three different types of cache ("F", "T" and "S") were applied to the BNC dialogue data – a fixed cache ("F"-type experiments), a cache of the current and the previous turn in the dialogue ("T"-type experiments) and a cache consisting of the current and the previous sentence ("S"-type experiments).

In each experiment, a unigram language model was constructed for the cache, and this interpolated with a simple trigram model, also trained on BNC dialogue material. For each experiment, a 10-fold cross-validation procedure (as described in section 4.6) was employed such that, in each rotation, 10% of the available data (approximately 770 000 words) was held back for testing. Of the remaining 90%, 300000 words were reserved for calculating the interpolation parameters and the remainder of the data (approximately 6 930 000 words) used for training the models. The optimal interpolation parameters for each case were computed as described in section 5.1 above. The average values of the parameters across the 10 rotations were calculated and the interpolated models re-applied to the 10 test datasets using these

weightings. Perplexity scores were computed for each of the resulting interpolated models. (It was not anticipated that the individual cache models would produce very meaningful perplexity scores in their own right, since they were typically constructed using rather small – and, in some cases, variable - amounts of data and were only based on unigram statistics.)

The perplexity scores obtained for these interpolated trigram-cache models were compared with perplexities both for the simple trigram model trained on dialogue, and for an "equivalent text" (TEQ) model trained on approximately the same amount of ordinary text material from the BNC. The "F" and "S" type cache experiments were also applied to the text-trained language models. The "T" type experiment is clearly not appropriate to ordinary text data where the concept of "dialogue turn" is not generally meaningful.

The "baseline" perplexity figures for the ordinary trigram models (averaged over 10 cross-validation rotations) were 185.97 for the dialogue-trained model and 532.94 for the TEQ model trained on the same amount of BNC text.

The summary results for these experiments are shown in table 5.1 below.

The results of these experiments are quite encouraging – with perplexity reductions of up to approximately 14% for dialogue and 27% for text data obtained – despite the cache model being very simple both conceptually and in its implementation. The interpolation weights for the cache components, whilst much smaller than 0.5 (implying that the hybrid model still relies more on information from its trigram component than from the cache model), are considerably larger than zero, showing that the cache is making a useful contribution.

Model	Aspect	Model Trained & Tested on Dialogue	"TEQ" Model Trained & Tested on Text
Simple trigram model only (baseline)	Average Perplexity (Perplexity Range)	186.0 (168.3 to 202.3)	532.9 (451.4 to 597.9)
Interpolated trigram & fixed-size (500 word) cache ("F" type experiment)	Average Perplexity (Perplexity Range)	160.4 (148.7 to 173.7)	389.0 (322.7 to 447.1)
	Perplexity Reduction w.r.t. Baseline	13.8 %	27.0 %
	Interpolation weight for cache	0.114	0.167
Interpolated trigram & previous turn as cache ("T" type experiment)	Average Perplexity (Perplexity Range)	165.9 (153.7 to 180.6)	Not Applicable
	Perplexity Reduction w.r.t. Baseline	10.8 %	Not Applicable
	Interpolation weight for cache	0.077	Not Applicable
Interpolated trigram & previous sentence as cache	Average Perplexity (Perplexity Range)	169.5 (156.0 to 184.9)	475.8 (397.2 to 535.9)
	Perplexity Reduction w.r.t. Baseline	8.9 %	10.7 %
	Interpolation weight for cache	0.060	0.057

Table 5.1 : Summary of results from experiments using cache models. The quoted perplexity values are across 10 cross-validation rotations, with the "average" values based on logprob scores weighted by dataset size.

It can be seen that, within this set of experiments, the fixed 500 word cache gives the best improvement over the baseline (trigram only) perplexity figure for both dialogue and text. It should be noted that the mean number of words per turn in the BNC dialogue material is 13 (the median number is just 5.5). This suggests that, even in dialogue material, there are significant re-uses of words at distances beyond just the most recent sentence or turn. However, bearing in mind that 80% of the dialogues (according to the definition we adopted in chapter 4) contain 55 words or less and 90% contain at most 112 words, use of a cache as big as 500 words may be unnecessary. This is investigated in some further experiments described below.

5.2.2 Variation of the Cache Size

Bearing in mind the above comment regarding the size of the fixed cache relative to typical lengths of turns and dialogues, a series of "F"-type experiments was carried out to investigate the effect of changing the size of a fixed cache on optimally interpolated trigram-cache models for "TEQ" and ordinary dialogue material respectively.

Summaries of the results obtained are shown in tables 5.2 and 5.3 below. As can be seen from these results, incorporating any cache-based component into the language models makes an improvement for both text and dialogue material from the BNC. Even a cache as small as 10 words makes a clear improvement to the perplexity in both situations when the resulting cache model interpolated with the ordinary trigram model. However, whereas increasing the size of the cache to as much as 1000 words continues to improve the perplexity of the models trained on tested on text material, in the case of the models trained and tested on dialogue data, the perplexity of the interpolated model is lowest for a cache size of around 300 words and very little improvement is obtained by increasing the cache size beyond 100 words. This may in part be due to the relatively short length, as noted above, of many of the BNC dialogues. Indeed, noting that Clarkson and Robinson (1997) used a mixture of text and spoken material from the BNC in their experiments may explain why they found

that a cache size of 500 words was optimal whereas this present experiment suggests that caches larger than that may give better perplexities for BNC text data.

Cache size (words)	Perplexity of Interpolated Model			Interpolation Parameters	
	Maximum	Minimum	Average	Trigram	Cache
0 (baseline)	202.3	168.3	186.0	1.000	0.000
10	190.2	159.3	174.6	0.961	0.039
20	183.7	154.0	168.5	0.939	0.061
50	177.5	149.8	163.1	0.914	0.086
100	175.1	148.5	161.1	0.901	0.099
200	174.1	148.3	160.3	0.892	0.108
300	173.8	148.3	160.3	0.889	0.111
400	173.7	148.6	160.3	0.887	0.113
500	173.7	148.7	160.4	0.886	0.114
1000	173.8	149.3	160.7	0.884	0.116

Table 5.2 : Comparison of perplexities and interpolation parameters of interpolated trigram-cache models trained and tested on dialogue material from the BNC for various different cache sizes. The Maximum, Minimum and Average perplexity values quoted are across 10 cross-validation rotations with the interpolation parameters fixed at the quoted values. Those quoted interpolation parameters are averages across 10 cross-validation rotations of the optimal values obtained by the EM algorithm for individual rotations.

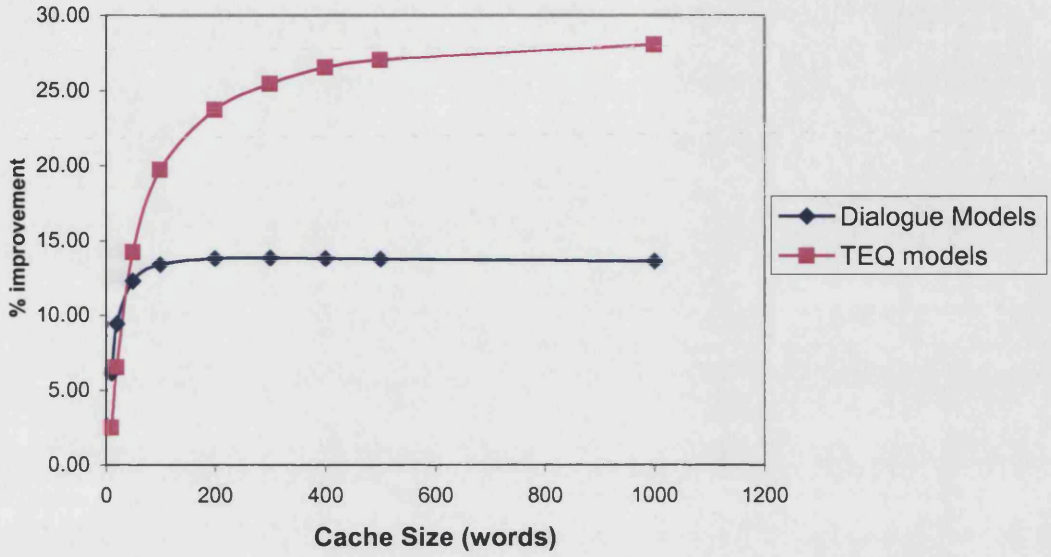
This trend is further investigated in terms of relative perplexity improvement over the baseline of the trigram model alone in figure 5.1 below. It can be seen that, in terms of relative perplexity improvement, small caches of just a few words give a greater benefit for the case of dialogue data than for ordinary text. Although the relative improvement obtained initially increases with increasing cache size for both dialogue and text, for dialogue it quickly reaches a plateau, whereas for text it continues to increase. Again, this may partly be due to the nature of the data in the BNC – some text samples contained therein are very long compared with the majority of the

dialogues. However, it may also be true that word repetitions – the feature essentially being exploited by cache models – are more significant at longer distances in text material than they are in dialogue, with the reverse being true for short-range repetitions.

Cache size (words)	Perplexity of Interpolated Model			Interpolation Parameters	
	Maximum	Minimum	Average	Trigram	Cache
0 (baseline)	597.1	487.1	532.9	1.000	0.000
10	583.2	441.1	519.6	0.988	0.012
20	560.6	422.7	498.2	0.969	0.031
50	515.7	386.6	457.0	0.930	0.070
100	485.3	360.1	427.8	0.897	0.103
200	462.6	339.1	406.5	0.867	0.133
300	453.7	330.3	397.4	0.851	0.149
400	449.4	326.0	391.6	0.841	0.159
500	447.1	322.7	389.0	0.833	0.167
1000	444.7	315.1	383.6	0.811	0.189

Table 5.3 : Comparison of perplexities and interpolation parameters of interpolated trigram-cache models trained and tested on "TEQ" material (a text sub-corpus of equivalent size to the dialogue data) from the BNC for various different cache sizes. The Maximum, Minimum and Average perplexity values quoted are across 10 cross-validation rotations with the interpolation parameters fixed at the quoted values. Those quoted interpolation parameters are averages across 10 cross-validation rotations of the optimal values obtained by the EM algorithm for individual rotations.

Figure 5.1 : Relative perplexity improvement of an interpolated trigram-cache model over a trigram model



For the interpolated trigram-cache models for dialogue, it was found that a fixed-size cache of just 20 words performed better (in terms of perplexity during testing) than the cache based on the current and previous sentence (experiment "S") and a fixed-size cache of 50 words or more out-performed the cache based on the current and previous turn (experiment "T"). This confirms that word repetitions (the basis of cache models) may be a significant factor in dialogue over a distance-scale longer than just successive turns. The continued improvement in perplexity up to cache sizes of around 100 words – comparable to the 90th percentile (112) for BNC dialogue length - suggests that such repetitions are relatively commonplace over the entire span of a dialogue.

5.3 Cache Experiments on DRT Data

As noted in section 5.1, it would be expected that particularly strong connections would hold between successive speaker turns of a dialogue. However, it is less clear whether it would be expected that the same words would necessarily be present with

high probability in consecutive speaker turns. To investigate this, and whether a cache model could be of significant value in a turn-prediction situation, e.g. for a spoken dialogue system, experiments were carried out where cache models were constructed separately : one for the first turns of pairs, the other for that part of the second turn which had already been processed. Although the cache window was limited to a maximum of 500 words, in practice this should not have imposed any restriction, due to the relatively short nature of the DRT turn pairs. Both caches were reset at the start of each new pair of turns.

The language models were trained on 90% of the available DRT material (423 000 turn pairs), with 5% (235 000 pairs) retained for each of development (computation of interpolation parameters) and evaluation (testing). A 10-fold cross-validation procedure was used, so that in each rotation a different 10% of the data was reserved for development and testing and the trigram and cache models trained on the other 90%.

As for the cache experiments on the ordinary dialogue material, the cache models were based on unigram statistics and were not considered of particular interest in their own right. However, the cache models for the second turns of the pairs were interpolated with the corresponding trigram model (trained on the same dataset) for second turns and a comparison of perplexity scores made with the baseline of the trigram model alone. Optimal interpolation parameters for each rotation were found using the EM algorithm, the average values of them calculated across the 10 rotations, the models re-applied to the separate datasets using these average values of the weightings and the revised perplexities evaluated. Average perplexity scores, based on logprob values weighted according to the exact sizes of the datasets used in each rotation, were calculated across the 10 rotations.

The results across the 10 rotations are summarised in table 5.4 below. In this case, unlike for the experiments on the ordinary dialogue material, no "equivalent text" experiment can be carried out since there is no natural equivalent of turn pairs in ordinary text.

	Trigram model for second turns alone	Interpolated trigram-cache model for second turns
Maximum perplexity for any one rotation	261.75	223.83
Minimum perplexity for any one rotation	145.58	134.19
Average perplexity (based on logprob scores) across 10 rotations	187.69	166.64

Table 5.4 : Summary comparison of perplexity scores across 10 cross-validation rotations for trigram only (baseline) and interpolated trigram-cache models. In both cases, these are for the second turns of pairs only. The optimal interpolation parameters were found to be 0.93387 for the trigram model and 0.06613 for the cache model. The interpolated model with these parameters gives an improvement in perplexity of 11.22% over the baseline of the trigram model alone.

The reduction in average perplexity obtained by including the cache component represents an improvement of 11.2% over the baseline figure, showing that incorporating this type of cache model into a language model for dialogue turn pairs can be useful. This is despite the DRT turn pairs being relatively short – a feature which might be expected to reduce the utility of a cache model. Intuitively, it might be expected that it would be difficult to obtain sufficiently reliable cache statistics for very short documents for the cache model to be of much use. On the other hand, however, the highly-varied nature of the BNC material might make it particularly suitable to benefit from short-term adaptability (Clarkson 1999, p 63), such as is offered by a cache model applied to short dialogue turns.

5.4 Qualitative Observations on Results from Cache-Based Models

In order to investigate the effect of using a cache model on individual turn pairs, a program was constructed to compare the probabilities given to individual dialogue turns according to a simple trigram model and according to a cache model alone. The turns showing the greatest ratio of cache model probability to trigram model

probability were output, and their content (together with the content of the immediately preceding turn) studied.

Many of the turns showing the greatest ratio of these model probabilities contained whole phrases "echoed" (or repeated) from the previous dialogue turn – probably to indicate agreement with the previous speaker, obtain clarification or indicate surprise in many cases. Some examples of such turn pairs are given in the table below, where the number preceding the text is the logarithm to base ten of the ratio of cache model probability to trigram model probability for the second turn of the pair. The text in square brackets is the preceding dialogue turn, whilst the turn of current interest follows. The + and - signs indicate that the word preceding it had probability at least 10% greater (or less, respectively) according to the cache model than to the trigram model.

197.878 [THE RAIN CAME DOWN AND THE FLOODS CAME UP AND THE HOUSE ON THE ROCKS SLID DOWN BUT THE FOOLISH MAN BUILT HIS HOUSE UPON THE SAND THE FOOLISH MAN BUILT HIS HOUSE UPON THE SAND THE FOOLISH MAN BUILT HIS HOUSE UPON THE SAND AND THE RAIN CAME TUMBLING DOWN] THE+ RAIN+ CAME+ DOWN+ AND THE+ FLOODS+ CAME+ UP- AND- THE+ HOUSE+ ON- THE- ROCKS+ SLID+ DOWN- BUT+ THE+ FOOLISH+ MAN+ BUILT+ HIS+ HOUSE+ UPON+ THE- SAND+ THE+ FOOLISH+ MAN+ BUILT+ HIS+ HOUSE+ UPON+ THE- SAND+ THE+ FOOLISH+ MAN+ BUILT+ HIS+ HOUSE+ UPON+ THE- SAND+ AND- THE+ RAIN+ CAME+ TUMBLING+ DOWN+

118.863 [SPLISH SPLASH AND THE RAIN CAME DOWN AND FLOODS CAME UP THE RAIN CAME DOWN AND THE FLOODS CAME UP THE RAIN CAME DOWN AND THE FLOODS CAME UP AND THE HOUSE ON HIS BAND FELL FLAT] AND+ THE+ RAIN+ CAME+ DOWN+ AND+ FLOODS+ CAME+ UP- THE+ RAIN+ CAME+ DOWN+ AND+ THE+ FLOODS+ CAME+ UP THE+ RAIN+ CAME+ DOWN+ AND+ THE+ FLOODS+ CAME+ UP AND+ THE+ HOUSE+ ON- HIS- BAND+ FELL+ FLAT+

101.600 [THE ROAD WAS BENDY AND TWISTY WITH LARGE SHADY TREES ON EITHER SORRY I'LL START AGAIN THE ROAD WAS BENDY AND TWISTY WITH LARGE SHADY TREES ON EITHER SIDE FORMING A BEAUTIFUL AVENUE] THE+ ROAD+ WAS+ BENDY+ AND+ TWISTY+ WITH+ SHADY+ TREES+ WITH+ LARGE+ SHADY+ TREES+ EITHER+ SIDE- FORMING+ A- LARGE+

83.118 [ALRIGHT YOU CAN GO LONGER TRULY SIR ALL THAT I LIVE BY IS WITH THE AWL HA HA HA I MEDDLE WITH NO TRADESMEN'S MATTERS NOR WOMEN'S MATTERS BUT WITH THE AWL] TRULY+ SIR+ ALL+ THAT- I+ LIVE+ BY+ IS+ WITH+ THE- AWL+ I- MEDDLE+ WITH+ NO+ TRADESMEN'S+ MATTERS+

70.712 [NO LISTEN I REALLY CAN'T MAKE THE APPOINTMENT MY SECRETARY WILL PHONE ON TUESDAY TO REARRANGE ANOTHER APPOINTMENT] I+ REALLY+ CAN'T MAKE+ THE- APPOINTMENT+ MY+ SECRETARY+ WILL+ MY+ SECRETARY+ WILL+ PHONE+ ON+ TUESDAY+ TO- REARRANGE+ ANOTHER+ APPOINTMENT+

63.445 [YOU NOT GUESS WHAT A A HORSE'S FAVOURITE TELEVISION PROGRAMME ARE IS] WHAT+ IS+ A+ HORSE'S+ FAVOURITE+ TELEVISION+ PROGRAMME+ WHAT+ IS+ A+ HORSE'S+ FAVOURITE+ TELEVISION+ PROGRAMME+

54.366 [OKAY ALLOTROPES OF CARBON A CARBONATE PLUS AN ACID GIVES] AN+ ACID+ CARBONATE+ PLUS+ AN+ ACID+

45.372 [I LAID A PENNY NO I SPENDED A BAKER SHOP AND ONE TOOK AWAY FIVE CURRANT BUNS IN A BAKER'S SHOP ONE WENT] NO+ FIVE+ CURRANT+ BUNS+ IN+ A+ BAKER'S+ SHOP+

44.889 [YES THANK YOU LISTEN THE SHOES ARE NOW REPAIRED ONE HEEL WAS WORN DOWN WHAT DID I JUST SAY] THE+ SHOES+ ARE+ NOW+ REPAIRED+ ONE+ HEEL+ WAS+ WORN+ DOWN+

42.863 [I'M SO GLAD THAT SHE'S MY LITTLE GIRL SHE'S SO GLAD SHE'S TELLING ALL THE WORLD THAT HER BABY BUYS HER THINGS YOU KNOW] SO+ GLAD+ THAT+ SHE'S+ MY+ LITTLE+ GIRL- SHE'S+ SO+ GLAD+ SHE'S+ TELLING+ ALL+ THE- WORLD+

42.403 [YEAH JUST ONE AND WE WANTED ONE PILAU AND ONE MUSHROOM ONE PILAU AND ONE MUSHROOM AND ONE MUSHROOM RICE] ONE+ PILAU+ RICE+ AND ONE+ MUSHROOM+

39.328 [ON WEDNESDAY I HAVE TO VISIT THE DENTIST I DO HOPE I WILL NOT NEED TO HAVE ANY FILLINGS WHAT DID I JUST SAY] ON+ WEDNESDAY+ I+ HAVE+ TO- VISIT+ THE- DENTIST+ I+ DO+ HOPE+ I+ WILL+ NOT- NEED+ ANY+ FILLINGS+

38.395 [COME ON YOU REDS HA COME ON] COME+ ON- YOU+ REDS+ COME+ ON+ YOU+ REDS+

36.489 [ONE FROM ELIZABETH AND ONE FROM RON] ONE+ FROM+ RON+ AND+ ONE+ FROM+ ELIZABETH+

36.433 [HARD RETURN AND SOFT RETURNS] HARD+ RETURN+ SOFT+ RETURN+

35.812 [TWO TWO AND ONE THAT WOULD MAKE FIVE AND THEN YOU NEED ANOTHER THREE WHY DON'T YOU HAVE THE LOOK FOR THE THREE P FIRST AND TICK THAT OFF AND THEN LOOK FOR THE FIVE P NEXT AND TICK THAT OFF] TICK+ FIVE+ P+ OFF+ TICK+ FIVE+ P+ OFF+

34.557 [JUST STICK TO WHAT YOU KNOW THERE'S CHICKEN SUPREME I SHOULD HAVE YOUR CHOW MEIN THEN OR CHICKEN AND MUSHROOMS OH NO THAT'S POT RICE THAT'S NOT POT NOODLE] I- KNOW- THERE'S+ POT+ RICE+ AND- POT+ NOODLE+ POT+

34.401 [WHERE'S TAB OH] TAB+ TAB+ TAB+

33.706 [THAT'LL BE ANOTHER FAMILY SAYING WON'T IT HAPPY BOILED EGG THAT WITH ALL THESE BOILED EGGS TO ROBERT] HAPPY+ BOILED+ EGG- HAPPY+ BOILED+ EGG-

32.922 [BUT THERE'S ONE SCORE WHICH UNAMBIGUOUSLY CALLS IT IN C MINOR ON THE AND IT STARTS IN C MINOR SO] IT+ IT+ STARTS+ IN+ C+ MINOR+ IT+ STARTS+

31.786 [SHIFT F SEVEN] SHIFT+ SHIFT+ F+ SEVEN+
31.085 [ONE ONE THIRD JUST WRITE DOWN ONE THIRD ADD ONE SIXTH] ONE+ THIRD+ ADD+ ONE+ SIXTH+
30.484 [COME ON LET'S HEAR DELLA SPEAK] DELLA+ DELLA+
30.209 [A THREE LEGGED CAT TIGER IT'S NAME'S TIGER IN N IT] A+ THREE+ LEGGED+ TIGER+
30.077 [YEAH IT WAS BEFORE DEANA WAS BORN] WAS+ IT- BEFORE+ DEANA+ WAS+ BORN+
29.536 [BEEN ON THE TRAIN TOOT TOOT BEEN ON THE TRAIN WHAT'S THE TRAIN SAY TRAIN SAY] TOOT+ TOOT+
28.648 [TAKE AWAY ONE TWELFTH OKAY JUST WRITE THAT DOWN THAT YOU'VE GOT TO TAKE AWAY THE ONE TWELFTH] ONE+ TWELFTH+ ONE+ TWELFTH+
28.480 [AND YOU'D BE RIGHT OKAY THE AVAILABILITY WE'VE LOOKED AT JOINT LIFE JOINT LIFE FIRST CLAIMS LIFE OF ANOTHERS SINGLE LIVES WHAT SORT OF BENEFIT WOULD CLAIM BE WRITTEN ON] SINGLE+ LIFE+ SINGLE+ LIFE+
27.447 [EAST HERTS YEAH] YEAH+ EAST+ HERTS+
27.426 [CLUB BAR LICENCE ALAS] CLUB+ BAR+ LICENCE+
26.836 [VERY WELL ON THAT NITRATES SULPHATES AND WHAT ELSE ANY OTHER HATES THAT YOU'VE HEARD OF] NITRATES+ SULPHATES+
26.644 [I MEAN AND I'D WORKED IN ONE IN OLDHAM AND THE DIFFERENCE BETWEEN THE ONE THAT WAS A FIFTY SHILLING TAILORS IN OLDHAM AND] FIFTY+ SHILLING+ TAILORS+
26.306 [IT WAS HER BROTHER WASN'T IT YES THEY WAS ALL BROTHERS WEREN'T THEY EDWARD AND THE KING AND THE DUKE OF KENT THEY'RE ALL BROTHERS] EDWARD+ THE- DUKE+ OF- KENT+ AND+ THE+ KING+ THEY+
26.105 [ONLY I WANT ONLY JUST GET ME A TIN OF HAIR LACQUER NORMAL HOLD FOR TINTED THAT'S ALL I WANT] NORMAL+ HOLD+ TINTED+
25.943 [IT WAS CREEPING SUBURBIA THEN IT WASN'T] CREEPING+ SUBURBIA+
25.869 [IT'S NATIONAL CROQUET DAY DID YOU REMEMBER THAT] NATIONAL+ CROQUET+ DAY+
25.018 [ISN'T IT IF THAT'S FIVE OR MORE THEN YOU INCREASE THAT BY ONE AND MAKE IT FIVE SO THE JOURNEY TIME IS ABOUT FIVE HOURS SO D FOUR WORK OUT THESE JOURNEY TIMES TO THE NEAREST HOUR BOMBAY TO PERTH AT ABOUT EIGHT HUNDRED AND FIFTY KILOMETRES AN HOUR] BOMBAY+ TO+ PERTH+ EIGHT+
23.826 [IT MUST HAVE BEEN GOOD IT MUST HAVE BEEN SOME SORT OF CARBONATE AND THE SALT THAT WAS FORMED WAS FROM THE HYDROCHLORIC ACID WAS CALCIUM CHLORIDE SO IT MUST HAVE BEEN] CALCIUM+ CARBONATE+
23.811 [SO YOU'D NAME THIS AS BUTANE IN OTHER WORDS YOU'RE SAYING IT'S A BUTANE CHAIN YOU TAKE OFF THE E YOU WILL ADD O L AND IF THERE ARE POSITIONAL ISOMERS POSSIBLE YOU HAVE TO INDICATE THE POSITION ONE O L BUTANE ONE L ONE O L BUTANE ONE O L] BUTANE+ ONE+ O+ L+

```
  23.110 [ IN FOURTEEN NINETY TWO YES BUT THAT'S BESIDE
THE POINT SO SHH PLEASE TANYA THE WATERS RETURNED FROM
OFF THE EARTH SO IT'S SAYING THE WATERS RAN OFF THE EARTH
SUBSIDED FROM ] WATERS+ SUBSIDED+
  22.523 [ LITTLE BO PEEP HAS LOST HER SHEEP AND DOESN'T
KNOW WHERE TO FIND THEM LEAVE THEM ALONE ] SHEEP+
DOESN'T+ WHERE+ TO+ FIND- THEM+
  22.093 [ WE'RE THE TRIGGER WE'VE GOT TO TRY AND
SCHEDULE THREE PICKUPS THREE BAGGINGS A DAY ] THREE+
PICKUPS+ A+ DAY+
  21.665 [ WELL PERSONALLY I MEAN MARGARET IMPRESSED ME
GREATLY BUT I THINK JOHN WAS INCLINED TO BE FULL OF HIS
OWN IMPORTANCE FOR ONE THING HE WAS EXCEPTIONALLY
DEMANDING AND MY FAIRLY LONG CONVERSATION WAS THAT HE
ALMOST LEAD ME TO BELIEVE THAT HE HAD GOT THE JOB BECAUSE
WHEN HE STARTED MAKING COMMENTS ABOUT PUTTING HIS
GRANDFATHER CLOCKS IN THE CHURCH THIRTY OF THEM AND
BUILDING A ] THIRTY+ GRANDFATHER+ CLOCKS+
  21.748 [ THERE'S ACTUAL STRAWBERRIES THERE ] ACTUAL+
STRAWBERRIES+
  21.611 [ BAR NINE IS BAR ONE AN OCTAVE LOWER HERE'S THE
BIT THAT'S IMPORTANT THIS IS BAR NINE AN OCTAVE LOWER
OKAY NOW THEN WILL YOU PLEASE COPY PRECISELY WHAT IS
THERE AT BAR TEN THE MUSIC YOU NEED ONE BAR OF MUSIC LINE
WITH THOSE BLOBS WHICH ARE THE NOTE HEADS IN EXACTLY THE
RIGHT PLACES JUST COPY WHAT'S IN THE BOOK ] COPY+ WHAT+
BAR+ TEN+
  21.429 [ CAN'T DO ANY MORE THAN THAT SO FRACTION OF A
CIRCLE FRACTION OF A CIRCLE THIS IS GOING TO BE ONE THIRD
TIMES THREE SIXTY OR WE COULD SAY NUMBER OF DEGREES HERE
FRACTION OF A CIRCLE AS NUMBER OF DEGREES SO THAT'S A
THIRD OF THREE SIXTY DEGREES AND THAT'S A THIRD OF THREE
SIXTY OKAY GOING TO BE A FIFTH OF THREE SIXTY AND TWO
FIFTEENTHS ] FIFTEENTHS+ OF+ THREE+ SIXTY+
  21.358 [ YEAH YEAH BILKO WAS IN IT WEREN'T HE OLD PHIL
SILVERS WAS IN IT ] YEAH- PHIL+ SILVERS+
  20.517 [ OH SHOULD OF GOT YOU SOME MORE FROMAGE FRAIS
MICHAEL OH I WAS LOOKING AT THEM AS WELL WASN'T I AND I
DIDN'T GET THEM YOU'LL HAVE TO HAVE A A WOBBLY A
STRAWBERRY WOBBLER ] I+ HAVE+ A- STRAWBERRY+ WOBBLER+ AS+
WELL-
  20.461 [ BUY A BOTTLE OF WHISKY AND ORDER UP A HAGGIS ]
WHISKY+ AND- HAGGIS+
```

Table 5.5 : Examples of turns much more probable according to cache model than to trigram model, with the previous turn shown in brackets. The initial number is the logarithm to base ten of the ratio of probabilities of the turn according to the cache model relative to that given by the trigram model.

It is difficult to say anything very specific about the content of these turn pairs – which clearly come from very varied sources. Some seem to be from recitations of biblical stories and nursery rhymes, others from what appears to be a drama workshop (on Shakespeare's "Julius Caesar") whilst others are extracts from pop songs (possibly where two singers have similar parts, but not synchronised). Some are from conservational sources, some from interaction between an adult (e.g. a parent) an a young child, others from instructional sources (chemistry lessons and IT tutorials) and others from situations where food is being ordered. However, the common feature is that one or more words (in many cases, relatively uncommon words) from the first turn of the pair appear again in the second turn of the pair. Although this may appear obvious, considering the nature of the cache model, we can note that what we might expect on the basis of such intuition does seem to be observed in practice. The probabilities of normally uncommon words which occur in the cache are enhanced. The probabilities of utterances of which many of the constituent words are in the cache are greatly enhanced – more so than if just one or two words from a long phrase appear in the cache. This suggests that a cache model could be highly beneficial in modelling dialogue – particularly in situations where words or phrases said by one speaker are immediately repeated by the other speaker – perhaps to indicate surprise, request clarification or just to allow the speaker some thinking time.

Some examples of repetitions of "phatic" utterances used for greetings or farewells were observed – although relatively few of these are highly ranked by ratio of cache probability to baseline model, presumably because such sequences are relatively common in dialogue and hence use of the cache does little to enhance their probability. e.g.

[BUBYE DADDY] BUBYE+ DADDY+ (test dataset 4)

[ALRIGHT CYNTH] ALRIGHT+ CYNTH+ (test dataset 4)

[BYE MAGS] BYE+ MAGS+ (test dataset 3)

In these cases, it would appear to be the repetition of the more unusual proper name (or the unusual spelling "bubye" rather than "bye-bye" or "bye bye") which causes the large probability enhancement.

There are also several instances of a student or pupil repeating part or all of a teacher's utterance, again probably for emphasis or requesting clarification or confirmation, as part of his/her reply. E.g. :

```
[ RIGHT AND THAT'S A GENERAL REACTION THAT HAPPENS WITH
VIRTUALLY ANY ACID AND ANY ALKALI ] ACID+ AND+ ANY+
ALKALI+

[ ONE ONE THIRD JUST WRITE DOWN ONE THIRD ADD ONE SIXTH ]
ONE+ THIRD+ ADD+ ONE+ SIXTH+

[ THE CURRENT IN THAT RESISTOR ] THE+ CURRENT+ IN+ THAT+
RESISTOR+

 [ EXACTLY IT'LL BE COS WE LOOKED AT THIS LAST WEEK AS
YOU SAID AT THE END WHAT HAPPENS DRIPPING ACID ONTO ONTO
CHIPS NOW THE THINGS TO KNOW ABOUT ACIDS BASES AND SALTS
A METAL PLUS AN ACID WHAT HAPPENS ] A+ METAL+ AND- PLUS+
AN+ ACID+

 [ CAN'T DO ANY MORE THAN THAT SO FRACTION OF A CIRCLE
FRACTION OF A CIRCLE THIS IS GOING TO BE ONE THIRD TIMES
THREE SIXTY OR WE COULD SAY NUMBER OF DEGREES HERE
FRACTION OF A CIRCLE AS NUMBER OF DEGREES SO THAT'S A
THIRD OF THREE SIXTY DEGREES AND THAT'S A THIRD OF THREE
SIXTY OKAY GOING TO BE A FIFTH OF THREE SIXTY AND TWO
FIFTEENTHS ] FIFTEENTHS+ OF+ THREE+ SIXTY+
```

5.5 Summary

The results presented and discussed in this chapter have shown that cache-based models have a useful role in supplementing trigram language models for both text and dialogue material – perhaps to a surprisingly good extent (giving perplexity reductions of up to about 14% for dialogue and 27% for text over a baseline of a trigram model alone) bearing in mind the simplicity of the model. For the data (from the BNC) used

in this study, it has been found that very small caches give a better relative improvement for dialogue data than for text data, but that the reverse is true for larger cache sizes. That may at least in part be due to the nature of the material of each type in the BNC – most dialogues in it are relatively short, whereas some texts are very long, which would account for significant re-occurrences of words at longer ranges in the text than in the dialogue data. However, there is insufficient evidence at present to say whether this would or would not generalise to cases where both longer texts and dialogues could be studied.

The relative success of models employing very small-sized caches (of just 30 or even 10 words) for dialogue material is probably due to repetitions of words – and even entire phrases – being common over a short scale in dialogue. This is particularly true across consecutive speaker turns, where a repetition of something said by the first speaker will often be used by the second speaker to obtain clarification or indicate surprise, or to confirm that an instruction has been understood correctly, e.g. :

[ONE FROM ELIZABETH AND ONE FROM RON] ONE+ FROM+ RON+ AND+ ONE+ FROM+ ELIZABETH+ (Confirmation)

[SHIFT F SEVEN] SHIFT+ SHIFT+ F+ SEVEN+ (Confirmation)

[A THREE LEGGED CAT TIGER IT'S NAME'S TIGER IN N IT] A+ THREE+ LEGGED+ TIGER+ (Surprise ?)

[YEAH IT WAS BEFORE DEANA WAS BORN] WAS+ IT- BEFORE+ DEANA+ WAS+ BORN+ (Query for clarification)

[ONLY I WANT ONLY JUST GET ME A TIN OF HAIR LACQUER NORMAL HOLD FOR TINTED THAT'S ALL I WANT] NORMAL+ HOLD+ TINTED+ (Confirmation)

[IT'S NATIONAL CROQUET DAY DID YOU REMEMBER THAT] NATIONAL+ CROQUET+ DAY+ (Surprise ?)

It was notable that the majority of turns showing the largest relative increases in probability due to use of the turn-based cache model over the probability given by the

simple trigram model were relatively short – or at least contained few words which did not appear in the previous (cached) turn. An explanation of this is that only the words within the turn which were held in the cache contribute to the enhancement of the turn's probability. Hence, the probabilities of turns containing a high proportion of cached words are going be enhanced more (in relative terms) than those of turns containing a small proportion of words from the cache.

Although it was found that use of a cache gave worthwhile improvements for cache sizes at the level of sentences or dialogue turns, the best performance for dialogue was obtained for a fixed-cache of about 300 words. However, on the evidence of the results in this chapter, it would appear that larger cache sizes could give better results for text material, indicating that word repetitions tend to occur at quite long distance scales in text material. This is not entirely surprising if individual text documents tend to have a particular theme so that words characteristic to that topic will re-occur, in addition to words which are generally very common being repeated. Although the same principle would be expected to be partly true for dialogue, themes and topics within the conversation may evolve in a way which makes such repetitions less common at longer distances. This is consistent with the findings of Purver, Ginzburg & Healey (2002), who found that requests for clarifications in dialogue tended to be separated from their source by just one speaker turn. (A separation of one turn accounted for 85% of such clarification requests, whilst 95% were separated from their source by 3 turns or less. A similar pattern was observed when the separations were measured in sentences.) Likewise, repetitions of words which occurred very recently in the conversation are very likely to occur in attempts to "ground" the situation in the sense discussed in chapter 3 (Traum & Allen 1992). Oviatt & Cohen (1991) and Cohen et al (1990) also found that clarifications, and requests for them, occurred very frequently in task-oriented dialogues, and accounted for a relatively high proportion of the verbal interaction in such conversations.

Chapter 6 Experiments Using Language Models Based on Trigger Pairs

6.1 Motivation & Overview

As noted in the previous chapter, and in chapter 2, the theme of a document or conversation may evolve gradually and some method of allowing the language model to adapt to the current theme is therefore desirable in speech technology applications. Trigger models – where pairs of words which are commonly found to occur near each other are employed to modify "baseline" probabilities based on N-gram models – are an alternative to cache models as a means of doing this. Trigger models are in a sense a generalisation of cache models – instead of just looking for repetitions of words in the recent history of the document or utterance, trigger models look for words which are known to tend to occur in close proximity to each other. It would seem likely that trigger models should prove useful in the statistical modelling of dialogue – we would expect there to be strong correlations – of the kind which trigger models should be able to exploit – between the lexical content of neighbouring sentences and turns of a dialogue. We would expect that such correlations might be particularly strong between successive dialogue turns and hence trigger models might be of great benefit in the modelling of the DRT data. Furthermore, as noted in chapter 4, the DRT dataset was constructed with the aim of investigating "structural" as opposed to "topic-dependent" features of dialogue. Would any evidence of such features, such as common referents for pronouns, be apparent from a trigger-based model trained and tested on DRT data ? In this section, practical aspects of the implementation of this trigger-based model, trained by the maximum entropy method, are discussed.

In order to construct a practical trigger model, it was necessary to consider a large number of potentially useful word pairs. In most experiments, these were words of intermediate frequency. Function words such as "and", "but", "the", "is" are too common and the fact that pairs of these (or one of these and a word which is just relatively common) occur in the recent history do not imply anything about the topic or content of the document under consideration. On the other hand, pairs of very uncommon words (as in the Brest – Litovsk example in Chapter 2) would occur so

rarely that they probably would be of very little practical use as triggers even if they were strongly correlated. The method recommended by Rosenfeld (1996) and adopted in this study is to initially propose a very large number of potential trigger pairs, where both words of each pair are "intermediate frequency words" in our lexicon – i.e. we exclude the extremely common function words and also extremely rare words. This results in a large set of possible triggers – $(V - E_C - E_U)^2$ pairs, where V is the total number of distinct words in our lexicon, and E_C and E_U are the number of words excluded for being too common and too uncommon respectively. (Note that we cannot necessarily assume symmetry within a trigger pair : a word A triggering a word B later on may not necessarily be equivalent to B triggering A.) For example, if we consider just 10 000 words – much fewer than the full lexicon - after excluding the very common and uncommon words, this would mean we would have a list of 100 million possible trigger pairs to consider. If the presence or absence of all these triggers within the recent history were to be included as variables in a language model, the computational demands (both in terms of memory and of processor time) would be extremely high, particularly during the training phase where the relative importance of the different variables is calculated. To reduce the effect of this problem, a much less computationally intensive pre-selection process is carried out on trigger pairs. Following Rosenfeld (1996), we retain only those pairs showing a relatively high mutual information I(A,B) over the set of available training data :

$$I(A,B) = P(A, B) \log(P(B \mid A) / P(B)) + P(A, B') \log(P(B' \mid A) / P(B'))$$
$$+ P(A', B) \log(P(B \mid A')/ P(B)) + P(A', B') \log(P(B' \mid A') / P(B'))$$

where the probabilities are empirical estimates calculated by taking ratios of the appropriate counts over the training data, and A' indicates the absence of A, etc. This enabled feature-based models to be constructed using a number of features (trigger pairs) which was small enough so that training the models was computationally tractable, but with the models still retaining those features most likely to be useful.

Once the pre-selection of trigger pairs had been completed, those still under consideration are incorporated into an exponential probability model :

$$P_\lambda(w \mid \underline{h}) = \frac{\exp\left(\sum_{i \in trigger\ pairs} \lambda_i f_i(\underline{h}, w) \right)}{Z_\lambda(\underline{h})}$$

where $\underline{h}$ is the recent history prior to the word w of current interest, f_i is a binary feature such that $f_i(\underline{h}, w) = 1$ if, in the list of trigger pairs still under consideration, the i^{th} trigger targets the word w and the triggering word is in $\underline{h}$, but otherwise $f_i(\underline{h}, w)$ is zero, λ_i is a weighting factor indicating the relative importance of the trigger f_i and Z_λ is a normalisation factor to ensure that the complete set of probabilities sum to 1 across the space of possible words w for a given $\underline{h}$:

$$Z_\lambda(\underline{h}) = \sum_w \exp\left(\sum_{i \in trigger\ pairs} \lambda_i f_i(\underline{h}, w) \right)$$

A buffer was used to store the "recent history" of the current document or dialogue, which was then compared with the current word to determine which (if any) of the binary features was "currently active". The weighting parameters $\{ \lambda_i \}$ were chosen using the Maximum Entropy method (see section 2.4) with respect to the training data, applying the "Generalised Iterative Scaling" algorithm (Darroch & Ratcliff 1972, Berger et al 1996, Rosenfeld 1996) to perform the required optimisations. Features were added to the model in the iterative manner described in section 2.4, choosing at each stage the new feature which gave the largest improvement in the objective function over the value for the optimal model with one feature fewer included. The iterative process was repeated until either the maximum number of features specified had been included in the model, or addition of any further features gave negligible improvement to the objective function (equivalent to maximising the entropy).

Note that if no trigger pairs are "active" for the current version of the "recent history", this model gives a default probability of $1/W$, where W is the number of distinct words under consideration (the vocabulary less any words which have been excluded). This aspect can give words unrealistic probabilities when no triggers are currently active. For example, consider the case of a very rare word which occurs just m times in the entire training corpus containing C words in total, where (say) $m < 10$.

Under a simple unigram model, based on the training dataset, this word should be assigned a probability m/C, and estimates of the same word according to bigram or trigram models would be of a similar magnitude. However, when no triggers are active and we are using the full vocabulary of V words, the same rare word would be assigned a probability of 1/V. In the case of this study, $C \approx 7 \times 10^6$, $V \approx 50000$, so if m = 7 (say) the probabilities for that rare word would be : P (unigram) $\approx 1 \times 10^{-6}$, but P (trigger) $\approx 2 \times 10^{-5}$, twenty times larger, despite no triggers for that word being active !

The resulting trigger model was then interpolated with a simple trigram model, the optimal interpolation parameters being computed using the *interp* program from the CMU Language Modelling Toolkit, which makes use of the EM algorithm (Dempster, Laird & Rubin 1977), applied to data reserved for this purpose.

In a similar manner to the experiments performed using cache models (described in Chapter 5), two distinct groups of experiments were carried out. The former, using ordinary dialogue data from the BNC, with a sample of ordinary text data of equal size (the "TEQ" dataset) used for comparison, investigated the relative merits of using a fixed-size widow, the current and immediately previous dialogue turns, and the current and immediately previous sentences for the "recent history". The other type of experiment concentrated on the "DRT" dataset – pairs of consecutive, relatively short dialogue turns.

6.2 Trigger Model Experiments on Ordinary Dialogue Data from the BNC

6.2.1 Experiments where the number of triggers per target word was restricted.

As noted above, three sets of experiments were carried out, employing different types of "recent history windows" when searching for words which might potentially be triggers for the word of current interest, to construct and evaluate trigger-pair based language models for dialogue material. A "restricted lexicon" was constructed by taking the full dialogue lexicon (approximately 50 000 words), ordered by decreasing

frequency of occurrence within the dialogue portion of the BNC, from which the 10 most common words were excluded. The 10 000 most common words in the remaining list were than taken to form the list of "words of intermediate frequency" which were considered to be the best candidates for either "triggering" or "target" words of trigger pairs. This yielded a set of 100 million potential trigger pairs, which were ranked for their likely utility by their mutual information with respect to the training dataset. From the training set of approximately 7 million words of dialogue transcription we found 4957 pairs with an average mutual information greater than 10^{-5} bits. From an equivalently sized quantity of text, 3209 pairs were found. Only 900 pairs were common to the two styles. A further restriction was imposed in order to maintain a reasonable range of words which were targets for triggers (otherwise, in principle, it would be possible for all the trigger pairs being used to target exactly the same word) – a maximum of 10 distinct triggers were permitted which targeted any given word. Pairs which met all these criteria for a given training set were stored in a special "appropriate triggers" list.

A sliding window was used to scan the "recent history" of the current dialogue, with respect to the word currently of interest. Any triggering words from pairs in the appropriate list found within the sliding window then made the corresponding pair "active" for the target word at that point in the document. (Once the window had moved forward to an extent that that triggering word was no longer in the window, the pair reverted to being "inactive".) Using the training dataset, an exponential probability model was constructed using the trigger pairs as features, as outlined in section 6.1 above. For each type of experiment, a 10-fold cross-validation procedure was followed, with a different 10% of the available data (approximately 770 000 words) being retained for testing in each rotation. Of the remaining 90% for each rotation (approximately 7 million words), 300 000 words were reserved as "development data" for calculating optimal parameters to be used for interpolating the exponential (trigger-based) model with an ordinary trigram model.

Three different types of sliding window were used : a "fixed" window of 500 words (or back as far as the start of the current dialogue, if fewer than 500 words before) in "F" type experiments, a window consisting of the previous dialogue turn plus the part of the current turn which had already been seen in "T" type experiments, and a

window comprising the previous sentence and the part of the current sentence already seen ("S" type experiments). For comparison, "F" and "S" type experiments were also performed on the "TEQ" dataset (a selection of text material from the BNC of the same size as the dialogue material) – "T" type experiments not being appropriate since there is no natural equivalent of dialogue turns for ordinary text. In all cases, the window was reset ("flushed") at the start of each new dialogue or text document and the two most recent words in the "history" were excluded from the window since use of them was already being made within the framework of the trigram model.

During each set of experiments (i.e. 10 cross-validation rotations for either dialogue or TEQ data, using the same type of window throughout), optimal interpolation parameters for combining the current trigger model with an appropriate simple trigram model (trained on dialogue or text, as appropriate) were computed using the EM algorithm with respect to the reserved 300 000 words of "development" data for each rotation. Average values of the interpolation parameters were calculated across the 10 rotations and the models for each rotation applied to the corresponding reserved test dataset with the interpolation parameters fixed at the average values. Perplexity scores were evaluated for each rotation and weighted average perplexity values computed, based on logprob scores for each, weighted by the size of the appropriate dataset used in that rotation.

A summary of the results obtained is given in table 6.1 below, with more detailed results for "F", "T" and "S" type experiments in tables 6.2, 6.3 and 6.4 respectively.

Model	Average Perplexity : Model Trained on Dialogue	Average Perplexity : Model Trained on 6.9 M words of Text
Simple trigram only (baseline)	186.0	532.9
Interpolated trigram-trigger model with fixed-size (500 word) window	183.0	493.2
Interpolated trigram-trigger model with previous turn as window	183.1	Not Applicable
Interpolated trigram-trigger model with previous sentence as window	182.5	492.8

Table 6.1 : Summary of results from experiments using trigger models. The quoted average perplexity values are based on logprob scores averaged over 10 cross-validation rotations. The number of triggers used in each experiment was not fixed, but was limited by each target word being permitted a maximum of 10 possible triggering words.

These results may appear slightly puzzling at first. Although the incorporation of the trigger model does give a modest improvement for all 3 types of window used, the best performance would appear to occur when the sentence-based window is employed. This would imply that the most useful correlations between words which can be exploited by a trigger model occur on a very short range – of the order of one or two sentences (or around 10 to 20 words, since the mean length of a sentence is about 8.7 words for the BNC dialogue material). On this basis, it might be expected that the turn-based window would outperform the relatively large fixed (500 word) window, since typical turns within the BNC dialogue material are of the order of 30 words long (median words per turn 11, mean 30.2, 90^{th} percentile 51) but this is not found to be the case.

	Model Trained on Dialogue	Model Trained on 6.9 M words of Text
Maximum perplexity for any one rotation.	197.3	558.0
Minimum perplexity for any one rotation	166.5	418.7
Average (based on logprob scores) perplexity across 10 rotations	183.0	493.2
% Improvement over baseline (ordinary trigram model)	1.60	7.46
Interpolation weighting for trigger-based component	0.0356	0.0736

Table 6.2 : Summary results across 10 cross-validation rotations for interpolated trigram-trigger models using a fixed-size window of 500 words ("F" type experiments). In each case, the baseline trigram model was trained on dialogue or text as appropriate for its application. Approximately 2150 triggers were used in each of these experiments.

However, it should be noted that the three different types of experiment did not all use the same number of triggers. Due to the nature of the process used for pre-selecting triggers from the list of those possible, particularly the restriction that there could be a maximum of 10 triggers for any given target word, the number of triggers incorporated into the models ranged from approximately 2150 for the fixed window ("F" type) experiments, through about 2800 for the turn-based window ("T" type) experiments, up to approximately 4200 for the sentence-based window ("S" type) experiments. This may account for the "S" window apparently performing better than either the larger "F" or "T" windows. A more controlled set of experiments, using the same number of triggers throughout, would clearly be of value here, and such a set of experiments is described in section 6.2.2 below.

	Model Trained on Dialogue
Maximum perplexity for any one rotation	197.6
Minimum perplexity for any one rotation	166.6
Average (based on logprob scores) perplexity across 10 rotations	183.1
% Improvement over baseline (ordinary trigram model)	1.53
Interpolation weighting for trigger-based component	0.0314

Table 6.3 : Summary results across 10 cross-validation rotations for interpolated trigram-trigger models using the current and the previous dialogue turn as the window (i.e. "T" type experiments). Note that dialogue turns are not a meaningful concept for the case of ordinary text data. Approximately 2800 triggers were used in each of these experiments.

In any case, the results of these experiments are rather disappointing. Since a trigger model, in general, exploits more general correlations between words than does a cache model, we would expect a trigger-based model to out-perform a cache model, with all other factors being equal. However, this was not found to be the case here. For the dialogue data, even a cache of just 10 words (see section 5.2.2) outperformed all of these trigger-based models, whereas for the TEQ data, it took a cache of 50 words (nevertheless, still quite a small size) to perform better than either of the appropriate trigger-based models. The reason for this is not fully understood, but may be linked to repetitions of words (the basis of cache models) tending to occur frequently even at very short separations in dialogue material. In contrast to this, correlations between any two distinct words may tend to be rather weak in dialogue and in text occur over a longer range in a manner consistent with the theme of the document evolving gradually. Although the ways in which the two types of models are constructed are somewhat different, a cache model is in some senses similar to a

trigger model where the triggering word is constrained to be the same as the target word, but where the number of potential trigger pairs incorporated into the model is exactly as large as the vocabulary size. This is much larger than the number of trigger pairs in any of the trigger-based models of this study (approximately 50000 in the former (cache) case compared with approximately 4200 in the latter (trigger) case).

	Model Trained on Dialogue	Model Trained on 6.9 M words of Text
Maximum perplexity for any one rotation.	197.0	557.6
Minimum perplexity for any one rotation	166.1	418.6
Average (based on logprob scores) perplexity across 10 rotations	182.5	492.8
% Improvement over baseline (ordinary trigram model)	1.86	7.54
Interpolation weighting for trigger-based component	0.0335	0.0690

Table 6.4 : Summary results across 10 cross-validation rotations for interpolated trigram-trigger models using the current and the previous sentence as the window (i.e. "S" type experiments). In each case, the baseline trigram model was trained on dialogue or text as appropriate for its application. Approximately 4200 triggers were used in each of these experiments.

6.2.2 Controlled Experiments Using a Fixed Number of Triggers

In order to study the effect of varying the number of trigger used in a rather more controlled way – the above experiments did not necessarily use the same number of trigger pairs in each case – a set of experiments was carried out for which the number of trigger pairs to be used in each case was predetermined. The results are shown in tables 6.5 – 6.9 below.

Using current and previous sentence as window :

Number of triggers used.	Dialogue models			"TEQ" Text models		
	Maximum Perplexity	Minimum Perplexity	Weighted Average	Maximum Perplexity	Minimum Perplexity	Weighted Average
100	197.619	166.547	183.128	558.969	419.334	493.440
200	197.519	166.479	183.059	558.778	419.250	493.322
500	197.401	166.399	182.957	558.587	419.068	493.186
1000	197.246	166.336	182.856	558.351	419.020	493.052
2000	197.154	166.254	182.730	558.085	418.858	492.907

Table 6.5 : Variation of perplexities of interpolated trigram-trigger models for dialogue and TEQ data with the number of trigger pairs incorporated into the model. Current and previous sentence used as the window.

Number of triggers used.	Dialogue models			"TEQ" Text models		
	Weighted Average Perplexity	Relative Perplexity Improvement	Interpolation Weight for Triggers	Weighted Average Perplexity	Relative Perplexity Improvement	Interpolation Weight for Triggers
0 (baseline)	185.997	-	0.00000	532.936	-	0.00000
100	183.128	1.54%	0.02827	493.440	7.41%	0.06533
200	183.059	1.58%	0.02902	493.322	7.43%	0.06589
500	182.957	1.63%	0.02997	493.186	7.46%	0.06692
1000	182.856	1.69%	0.03085	493.052	7.48%	0.06749
2000	182.730	1.76%	0.03183	492.907	7.51%	0.06809

Table 6.6 : Improvement of perplexities and variation of interpolation weights of interpolated trigram-trigger models for dialogue and TEQ data with the number of trigger pairs incorporated into the model. Current and previous sentence used as the window.

As can be seen from tables 6.5 and 6.6 above, there is only a very weak improvement in the reduction in perplexity over the baseline trigram model as the number of triggers incorporated in the model is increased. The decrease in perplexity compared with the trigram model is particularly disappointing for the dialogue data. It is believed that this is largely due to the relatively short length of typical sentences in the dialogue data, resulting in very few triggers, on average, being active at any one time. The longer average length of sentences in the text data would then explain why the trigger models work better when applied to ordinary text.

Using the current and previous turn as the window (not appropriate for text data) :

Number of triggers used	Interpolated Trigram-Trigger Models for Dialogue Data				
	Maximum Perplexity	Minimum Perplexity	Weighted Average Perplexity	Relative Perplexity Improvement over baseline	Mean Interpolation Parameter for Triggers
100	197.684	166.602	183.181	1.51%	0.02820
200	197.668	166.603	183.171	1.52%	0.02867
500	197.658	166.619	183.175	1.52%	0.02917
1000	197.654	166.649	183.195	1.51%	0.02971
2000	197.606	166.651	183.168	1.52%	0.03077

Table 6.7 : Variation of perplexities and interpolation weights of interpolated trigram trigger models for dialogue data with the number of trigger pairs included in the model. Current and previous turn used as the window.

The results for this case (shown in table 6.7) are particularly disappointing. It is not clear why use of the turn-based window should give weaker results than the (on average, shorter) sentence-based window. The change in relative perplexity improvement over baseline as the number of triggers used is increased is negligible.

Using a fixed window of the last 500 words :

Number of triggers used.	Dialogue models			"TEQ" Text models		
	Maximum Perplexity	Minimum Perplexity	Weighted Average	Maximum Perplexity	Minimum Perplexity	Weighted Average
100	197.661	166.615	183.196	558.876	419.558	493.505
200	197.639	166.589	183.176	558.647	419.142	493.338
500	197.529	166.545	183.116	558.344	418.951	493.190
1000	197.411	166.518	183.074	558.097	418.77	493.063
2000	197.318	166.497	183.033	558.079	418.713	493.192

Table 6.8 : Variation of perplexities of interpolated trigram-trigger models for dialogue and TEQ data with the number of trigger pairs incorporated into the model. Fixed window of the previous 500 words.

Number of triggers used.	Dialogue models			"TEQ" Text models		
	Weighted Average Perplexity	Relative Perplexity Improvement	Interpolation Weight for Triggers	Weighted Average Perplexity	Relative Perplexity Improvement	Interpolation Weight for Triggers
0 (baseline)	185.997	-	0.00000	532.936	-	0.00000
100	183.196	1.51%	0.02849	493.505	7.40%	0.06574
200	183.176	1.52%	0.02918	493.338	7.43%	0.06736
500	183.116	1.55%	0.03120	493.190	7.46%	0.06927
1000	183.074	1.57%	0.03330	493.063	7.48%	0.07118
2000	183.033	1.59%	0.03533	493.192	7.46%	0.07314

Table 6.9 : Improvement of perplexities and variation of interpolation weights of interpolated trigram-trigger models for dialogue and TEQ data with the number of trigger pairs incorporated into the model.

Once again, in each case shown in tables 6.8 and 6.9, an improvement over the perplexity of the baseline trigram model is obtained when it is interpolated with the trigger model. However, little further improvement is obtained by increasing the number of triggers in the model above 100. The trigger model has a much bigger effect for the text data than for the dialogue data. However, this longer window (500 words) does not give better results than the sentence-based window, suggesting that such dependences are very short-ranged.

6.3 Trigger Model Experiments on DRT Data

As noted in section 6.1 above, and in chapter 5, it would be expected that very strong correlations at a lexical level would occur between successive turns within a dialogue and that some such correlations would relate to the nature and structure of dialogue rather than being topic-specific. For example, it would be expected that the turn following a "polar " question would have a high probability of containing one of the words "yes", "no" or "probably", or that the turn after any type of question might include the phrase "I don't know". The framework of trigger models allows such issues to be investigated in a more general way than with cache models, which only look for repetitions of the same words over a relatively short scale. Trigger models allow not only this (cache models effectively make use of words which are "self-triggers") but also take consideration of pairs of words which, during training, have tended to co-occur relatively close together. As discussed in section 5.3, better modelling of successive dialogue turns could be of benefit in speech technology applications such as dialogue systems. To investigate how useful trigger models could be in such a context, experiments were carried out on the DRT (Dialogue Reduced Turn) dataset of pairs of successive turns from the dialogue material in the BNC where the total length of the pair does not exceed 200 words.

A sliding window containing a maximum of 500 words previous to the current target word, either in the current turn or the previous turn, was chosen as the word-trigger history, with the window being reset after every pair of dialogue turns. The history did not include the two words immediately previous to the target (as these would form part of the trigram model with which the trigger model would later be interpolated).

As suggested in section 6.1, when searching for potentially useful trigger pairs, only intermediate frequency words were considered. The arbitrary criterion applied was to use (for the purposes of the trigger model) a lexicon of only 10000 words which consisted of the most commonly-found words in the dialogue material, excluding the 50 most common, as for the experiments described in section 6.2. The pre-selection process for triggers was also applied in the same way as for the experiments using ordinary dialogue data.

From the list of potential trigger pairs, a criterion of a maximum of 10 triggering words per target word was imposed in order that the triggers to be used targeted a wide selection of words. The 2 800 triggers both satisfying this constraint and showing the highest mutual information over the training data were used to construct a exponential probability model with 2 800 parameters evaluated from training data using the Maximum Entropy framework with the Generalised Iterative Scaling (GIS) algorithm as discussed in section 6.1. The resulting exponential model was interpolated with the baseline trigram model for the content of second turns of pairs using the EM algorithm and a reserved set of data.

As for the other experiments, a 10-fold cross validation process was applied. In a similar way to the DRT experiments using cache-based models (see section 5.3), during each rotation, 90% of the available data was used for training the models, 5 % reserved for calculating the interpolation parameters and 5 % retained for testing. The optimal interpolation parameters were calculated for each rotation, the values averaged across the 10 rotations, then the models re-applied to test data with the interpolation parameters fixed at these average values. The perplexity of each model, with respect to its appropriate test dataset, and the weighted average perplexity across the 10 rotations (based on logprob scores weighted according to the size of the dataset used in each case) computed. These values were compared with those for the baseline of the corresponding ordinary trigram model trained on DRT data. The results are summarised in table 6.10 below.

	Trigram model for second turns alone	Interpolated trigram-trigger model for second turns
Maximum perplexity for any one rotation	261.75	250.21
Minimum perplexity for any one rotation	145.58	143.70
Average perplexity (based on logprob scores) across 10 rotations	187.69	182.81

Table 6.10: Summary comparison of perplexity scores across 10 cross-validation rotations for trigram only (baseline) and interpolated trigram-trigger models where the trigger component incorporates 2800 trigger pairs. In both cases, these are for the second turns of pairs only, taken from the DRT dataset. The optimal interpolation parameters were found to be 0.967573 for the trigram model and 0.032427 for the trigger model. The interpolated model with these parameters gives an improvement in perplexity of 2.60% over the baseline of the trigram model alone.

The improvement of the interpolated model over the simple trigram model is rather disappointing, particularly in comparison with the larger improvement found when using an interpolated trigram-cache model on the same datasets and bearing in mind that the trigger model exploits more general correlations between words than the cache model does. Although the number of triggers used in this experiment is smaller than the number stated by Rosenfeld (1994, 1996) in his experiments on text material for the Wall Street Journal corpus, the number used here would appear to be close to the limit which is feasible for the computations to be performed in a reasonable time on the best computer hardware available locally (Pentium 4, 2.0GHz PC with 512Mbytes of RAM). Rosenfeld (1994, section 5.7) has discussed the issue of the computational resources required by this type of model, exploring options including use of parallel computer architectures or distributed computing and streamlining the computational algorithms used.

6.4 Comparison of "Best" Trigger Pairs for Dialogue, Text and DRT Data

It is interesting to compare the highest-ranking (in terms of mutual information) trigger pairs for each of the ordinary dialogues. This could yield insight into the different nature of words associations between these three distinct forms of material within the BNC (although, of course, the DRT data is a subset of the ordinary dialogue material). In the case of dialogue and DRT data, this could be related to the psycholinguistic concept of "priming" (e.g. Bodner & Masson 2003) and the semantic idea of how participants "ground" the conversation (Traum & Allen 1992), as discussed in Chapter 2.

Lists of the top few of these "best" triggers for each type of material are displayed in tables 6.11, 6.12 and 6.13 below

For the TEQ dataset using the 50000 word vocabulary, the top of the list of trigger pairs was dominated by relatively unusual trigger words (including a large number of proper names), mostly targeting either the word "award" or the word "zero". Although this shows that such pairs of words (as shown in table 6.11 below) tend to either occur together or not at all, it also shows the danger of basing a model on such "unrestricted" choices of trigger pairs. Although the appearance of words from such pairs of words may be strongly correlated with the appearance of the other word of that pair, if the words themselves are rather uncommon then will tend to be of very limited value - as in Rosenfeld's Brest-Litovsk example (Rosenfeld 1996) – when incorporated into a trigger model for predicting the next work in a sequence within text. Furthermore, this observation also illustrates the sensitivity of a model to the material it is trained on in comparison to the material to which it is to be applied, also noted by Rosenfeld (2000b). The trigger pairs noted here appear to have their origin in reports of legal proceedings (e.g. BNC files FBS to FE3) or reports from grant-awarding bodies (e.g. ESRC, SSRC). Such trigger pairs – with the probable exception of those involving proper names – might be extremely useful within a restricted domain, but are very unlikely to be beneficial to the modelling of language in a more general context.

The set of triggers listed in table 6.11a proves an interesting contrast with the other cases – even the situation where the best trigger pairs for the TEQ text data are being investigated using the restricted vocabulary (the 10000 most common words, excluding the 10 most common) yielded much less unusual triggers. In the latter case (table 6.11b), the top trigger pairs just comprised very common words with no particularly interesting linguistic properties. However, scrutinising a longer version of this list shows some trigger pairs where the triggering and target words are in some way connected. For example, the word "her" triggering "she" comes 35th in the list, whilst "she" triggering "her" comes 56th. It is not surprising that the third person singular subject pronoun tends to be associated with the presence of the third person singular possessive or object pronoun in relatively close proximity. Similarly, "him" triggering "his" comes 83rd in this list. (Note that "he" cannot appear in this set of triggers since it is one of the most common words in the BNC text material). The observation that "nineteen" features relatively strongly as both a trigger (104th triggering itself, 154th, 183rd in the list of pairs) and a target (129th, 144th, 199th and 238th in the list of pairs) word – often paired with "by", "from" or "as" - is probably due to its use in dates relating to years of the twentieth century. However, it is difficult to infer very much more from this set of triggers.

The corresponding lists for trigger pairs for BNC dialogue data (tables 6.12a, 6.12b) are more intriguing. There is a surprisingly high incidence of pairs of mathematical words near the top. These probably originate from files within the BNC data recorded during maths classes or tutorials – for example, files G61, GYP, GYX, J91, KND, KNE. Words of such pairs (e.g. "equals" and "X", "plus" and "minus") may be expected to be strongly correlated in their presence or absence, and would perhaps be extremely useful in predicting word sequences used during a maths class ! However, they are not expected to be very typical of everyday British English conversation.

M.I.	Trigger	Target	N(A,B)	N(A,B')	N(A',B)	N(A',B')
0.00268031	OMBUDSMEN	AWARD	4330	211253	324	7524759
0.00267687	SUDIES	AWARD	4298	202647	356	7533365
0.00266773	CARLEN	AWARD	4288	202088	366	7533924
0.00265628	VICTIMISATION	AWARD	4324	215858	330	7520154
0.00265611	HOLDAWAY	AWARD	4276	201603	378	7534409
0.00265409	LEVI	AWARD	4326	217072	328	7518940
0.00263757	CRIMINOLOGICAL	AWARD	4333	223841	321	7522823
0.00256208	BARRISTERS	AWARD	4304	236256	350	7499756
0.00254865	MCBARNET	AWARD	4300	238953	354	7497059
0.00254365	HOSPITALISATION	AWARD	4232	218744	422	7517268
0.00251639	PROBATIONERS	AWARD	4274	240089	380	7495923
0.00251298	INCARCERATION	AWARD	4305	251693	349	7484319
0.00251237	SUDIES	ZERO	6112	200833	6788	7526933
0.00251190	ILO	AWARD	4214	222240	440	7513772
0.00250365	CARLEN	ZERO	6094	200282	6806	7527484
0.00250280	BOSTOCK	AWARD	4105	193286	549	7542726
0.00250220	SSRC	AWARD	4105	193440	549	7542572
0.00249965	MOFFAT	AWARD	4230	230817	424	7505195
0.00249717	CORPORATIST	AWARD	4175	214619	479	7521393
0.00249652	HOLDAWAY	ZERO	6079	199800	6821	7527966
0.00249641	OMBUDSMEN	ZERO	6158	209425	6742	7518341
0.00248547	MAILED	AWARD	4257	243941	397	7492071
0.00248375	ESRC	AWARD	4298	258637	356	7477375
0.00247444	DETERMINISTIC	AWARD	4304	263858	350	7472154
0.00247358	KEELE	AWARD	4292	259834	362	7476178
0.00246825	PROTECTIONS	AWARD	4207	232857	447	7503155
0.00246614	VICTIMISATION	ZERO	6148	214034	6752	7513732
0.00246252	LEVI	ZERO	6152	215246	6748	7512520

Table 6.11a : The 28 "best" trigger pairs (according to their mutual information scores) for text data in the BNC, together with their occurrence and non-occurrence statistics. These were obtained using the top 50000 words in the text vocabulary and the TEQ dataset (of equivalent size to the BNC dialogue dataset) and a 500 word window. The restriction limiting the number of triggers per target word to 10 has been removed in this case.

M.I.	Trigger	Target	N(A,B)	N(A,B')	N(A',B)	N(A',B')
0.00039625	AS	AS	40317	3398253	3036	748200554
0.00034303	BY	AS	36196	3136727	7157	748462080
0.00032920	FROM	AS	35075	3080369	8278	748518438
0.00030448	AN	AS	32742	2817649	10611	748781158
0.00030117	WHICH	AS	32262	2690823	11091	748907984
0.00027952	AS	BY	28844	3409726	2981	748200609
0.00027921	OR	AS	30311	2570131	13042	749028676
0.00027665	BY	BY	28345	3144578	3480	748465757
0.00027213	ARE	AS	29776	2589444	13577	749009363
0.00026108	ARE	ARE	25836	2593384	3319	749019621
0.00025525	AS	ARE	26370	3412200	2785	748200805
0.00025223	HIS	HIS	23459	1909467	2675	749706559
0.00025097	HAD	HAD	24016	2175787	3032	749439325
0.00024867	FROM	BY	26276	3089168	5549	748521167
0.00023866	AS	HAD	24591	3413979	2457	748201133
0.00023858	BEEN	AS	26452	2278407	16901	749320400
0.00023544	AS	FROM	24415	3414155	2775	748200815
0.00023305	MORE	AS	25778	2165353	17575	749433454
0.00023279	BY	ARE	24502	3148421	4653	748464584
0.00023128	THEIR	AS	25669	2186497	17684	749412310
0.00022920	AS	HIS	23666	3414904	2468	748201122
0.00022894	WHICH	BY	24296	2698789	7529	748911546
0.00022842	AN	BY	24420	2825971	7405	748784364
0.00022688	TWO	AS	25495	2294271	17858	749304536
0.00022346	FROM	FROM	23253	3092191	3937	748522779
0.00022267	FROM	ARE	23691	3091753	5464	748521252
0.00022241	HAD	AS	24898	2174905	18455	749423902
0.00021752	BY	FROM	22882	3150041	4308	748464929

Table 6.11b : The 28 "best" trigger pairs (according to their mutual information scores) for text data in the BNC, together with their occurrence and non-occurrence statistics. These were obtained using a restricted text vocabulary of 10000 words (excluding the 10 most common and all uncommon words) and the TEQ dataset (of equivalent size to the BNC dialogue dataset) and a 500 word window.

This again illustrates the sensitivity of language models to the data on which they are trained relative to the material to which they are to be applied (Rosenfeld 2000b). If these mathematical terms are ignored, the majority of the highest-ranked trigger pairs are composed of words which would be expected to occur commonly in dialogue. Some are colloquial words such as "okay" and "aye", whilst others are essentially transcriptions of noises indicating vague agreement or puzzlement ("mm", "mhm"), sometimes known as "backchannels". Such backchannels are probably of significant value when the participants in the conversation are trying to achieve "grounding" of it (Traum & Allen 1992) - they can provide useful feedback to the other speaker(s) regarding to what extent the utterer of the backchannel understands what the conversation is about and whether any ambiguities need to be resolved.

Again, we see that some closely-related common words appear as highly-ranked trigger pairs – for example "she" and "her" act as triggers for each other, whilst "she's" acts as a trigger for both "she" and itself. In the case where the unrestricted dialogue vocabulary was used, "him" and "he's" and "his" all acted as triggers for "he". The presence of the pair "bracelet" and "pounds", in addition to several other lower-ranked pairs relating to jewellery and money, suggest that a significant part of the training dataset related to such transactions, where prices in pounds would indeed be correlated with names of items of jewellery. Such cases are examples which might be predicted from psycholinguistic studies of semantic "priming" (Bodner & Masson 2003, Holcomb 1993, Meyer & Schvaneveldt 1971).

The trigger pairs obtained from the DRT dataset when the triggering word was restricted to being in the first turn of the pair, with the target word in the second turn, showed similar trends to the set of triggers obtained for the ordinary dialogue data with a fixed window. However, words of greeting (e.g. "hello") and farewell (e.g. "bye") also featured prominently, as might be expected in successive turns. (See table 6.13 below.)

M.I.	Trigger	Target	N(A,B)	N(A,B')	N(A',B)	N(A',B')
0.00207572	SHE	SHE	17540	1821546	7320	4416396
0.00173906	HE	HE	28972	2619745	10947	3603138
0.00108820	HER	SHE	13264	1527158	11596	4710784
0.00101523	SHE'S	SHE	10866	1115127	13994	5122815
0.00099989	THE	THE	204256	4644248	36416	1377882
0.00082669	MINUS	MINUS	979	123248	69	6138506
0.00068960	X	X	1036	208322	194	6053250
0.00067906	SHE	HER	7297	1831789	4070	4419646
0.00067021	THEY	THEY	32852	3302866	13740	2913344
0.00066776	HER	HER	6614	1533808	4753	4717627
0.00063701	AND	AND	130410	4490373	29469	1612550
0.00062381	HIM	HE	18564	1803675	21355	4419208
0.00061367	HE'S	HE	18122	1750770	21797	4472113
0.00061297	OKAY	OKAY	5056	1424519	3393	4829834
0.00061170	WAS	WAS	35701	3444570	14332	2768199
0.00060457	MM	MM	15544	2322105	10615	3914538
0.00059451	OF	OF	76635	4164888	21609	1999670
0.00059253	SAID	SAID	13614	2523101	7305	3718782
0.00058883	WE	WE	35423	3434651	14519	2778209
0.00056482	SQUARED	X	735	72333	495	6189239
0.00053343	HUNDRED	HUNDRED	3656	1332236	2183	4924727
0.00051646	CHAIRMAN	YEAH	492	345881	65022	5851407
0.00051547	X	MINUS	818	208540	230	6053214
0.00050906	HIS	HE	17629	1770381	22290	4452502
0.00049912	SHE'S	SHE'S	3789	1122204	3473	5133336
0.00049620	COMMITTEE	YEAH	476	333336	65038	5863952
0.00049206	SQUARED	MINUS	636	72432	412	6189322
0.00049169	PLUS	MINUS	937	443222	111	5818532

Table 6.12a The 28 "best" trigger pairs (according to their mutual information scores) for ordinary dialogue data in the BNC, together with their occurrence and non-occurrence statistics. These were obtained using the full 50000 word dialogue vocabulary and the BNC dialogue dataset with a fixed-size 500 word window.

M.I.	Trigger	Target	N(A,B)	N(A,B')	N(A',B)	N(A',B')
0.00387373	SHE	SHE	17540	975172	7320	2399917
0.00203831	HER	SHE	13264	817708	11596	2557381
0.00188375	SHE'S	SHE	10866	599385	13994	2775704
0.00151116	MINUS	MINUS	979	68592	69	3330309
0.00126669	SHE	HER	7297	985415	4070	2403167
0.00126344	X	X	1036	114633	194	3284086
0.00124601	HER	HER	6614	824358	4753	2564224
0.00116826	MM	MM	15544	1237499	10615	2136291
0.00112575	OKAY	OKAY	5056	773261	3393	2618239
0.00111356	SAID	SAID	13614	1356486	7305	2022544
0.00103069	SQUARED	X	735	40356	495	3358363
0.00097307	HUNDRED	HUNDRED	3656	727453	2183	2666657
0.00094400	X	MINUS	818	114851	230	3284050
0.00092276	SHE'S	SHE'S	3789	606462	3473	2786225
0.00090630	MHM	MHM	2742	492381	2362	2902464
0.00090160	PLUS	MINUS	937	242909	111	3155992
0.00089785	SQUARED	MINUS	636	40455	412	3358446
0.00087789	MULTIPLY	MINUS	668	55153	380	3343748
0.00087563	AYE	AYE	1544	239034	1548	3157823
0.00086167	MINUS	X	732	68839	498	3329880
0.00085238	FUCKING	FUCKING	929	98049	960	3300011
0.00082783	MULTIPLYING	MINUS	560	28574	488	3370327
0.00082618	NOUGHT	MINUS	711	87850	337	3311051
0.00080873	EQUALS	X	731	81455	499	3317264
0.00080142	SHE'S	HER	4864	605387	6503	2783195
0.00078169	EQUALS	MINUS	675	81511	373	3317390
0.00077095	BRACELET	POUNDS	525	6646	2758	3390020
0.00075161	USED	USED	3799	846708	2477	2546965

Table 6.12b The 27 "best" trigger pairs (according to their mutual information scores) for ordinary dialogue data in the BNC, together with their occurrence and non-occurrence statistics. These were obtained using the restricted dialogue vocabulary (the most common 10000 words after the most common 10 had been excluded) and the BNC dialogue dataset with a fixed-size 500 word window.

Although in the case of the DRT data, the majority of the top-ranked amongst these "best" trigger pairs are between rather common words, some interesting pairs do appear. Some of these are due to the exact content of the training material and would not necessarily be expected to be typical for general English dialogue. For example, in the list of trigger pairs for ordinary dialogue material, pairs such as "bracelet – pounds", "necklace – pounds", "carat – pounds" feature relatively prominently. Such pairs are notable if the conversation is concerned with jewellery, but would not be expected to be particularly useful in more general circumstances. Such sensitivity to the content of a relatively small part of the training material suggests that, despite the efforts of the compilers of the BNC, the corpus is not "correctly balanced" with respect to the breadth and distribution of "typical" modern British English dialogue. It would appear that, unintentionally, certain topics (or even more restricted areas of conversation) – such as jewellery shopping or coverage of a particular single football match – have been given a disproportionate amount of coverage within the BNC relative to their relevance to everyday conversation.

However, attempting to compile a large "well-balanced" corpus would not necessarily be easy, and collecting genuinely spontaneous conversation in the street or over the telephone may raise ethical issues regarding the consent and civil liberties of the participants. Should material be collected or recorded prior to the participants giving their consent ? If so, it could be considered as infringing the privacy of the participants. However, on the other hand, receiving consent before the recording is made could influence what the participants say and how they say it – making the conversation less genuinely spontaneous, since the speakers are conscious of being recorded or monitored.

M.I.	Trigger	Target	N(A,B)	N(A,B')	N(A',B)	N(A',B')
0.00903856	SHE	SHE	2594	11926	11206	389477
0.00897929	HE	HE	3795	17512	16393	377503
0.00393853	THEY	THEY	3312	21331	20038	370522
0.00380156	BYE	BYE	245	618	467	413873
0.00371063	MM	AND	5068	13736	54487	341912
0.00342949	WAS	WAS	3531	23153	21817	366702
0.00334803	AND	MM	5395	58978	14251	336579
0.00290325	MM	THE	5935	12869	74359	322040
0.00278629	ONE	ONE	2592	19564	18554	374493
0.00276660	FIVE	FIVE	555	4560	4272	405816
0.00258297	THREE	THREE	633	5646	5311	403613
0.00240274	TWO	TWO	921	8770	8272	397240
0.00238876	SHE'S	SHE	822	4416	12978	396987
0.00237265	SHE	SHE'S	814	13706	4112	396571
0.00236642	HUNDRED	HUNDRED	338	2603	2432	409830
0.00236247	FOUR	FOUR	448	3916	3698	407141
0.00236106	HE'S	HE'S	750	7246	6755	400452
0.00231743	SHE	HER	992	13528	6333	394350
0.00222579	HE	HE'S	1220	20087	6285	387611
0.00220836	HER	SHE	980	6846	12820	394557
0.00220282	MM	A	4505	14299	55278	341121
0.00215342	MM	TO	4373	14431	53441	342958
0.00213374	TWENTY	TWENTY	353	3079	2919	408852
0.00210480	HELLO	HELLO	187	949	996	413071
0.00209474	MINUS	MINUS	117	267	263	414556
0.00208920	HE'S	HE	1198	6798	18990	388217
0.00206221	SIX	SIX	331	2923	2725	409224
0.00205261	YEAH	THE	12873	38380	67421	296529

Table 6.13 The 28 "best" trigger pairs (according to their mutual information scores) for DRT data in the BNC, together with their occurrence and non-occurrence statistics. These were obtained using the full 50000 word dialogue vocabulary and the BNC dialogue dataset, strictly using only the previous turn as window.

6.5 What Kind of Turn Pairs Benefit Most from the Use of a Trigger Model ?

In a similar manner to the study carried out to study the effect of the cache model on individual dialogue turns (see section 5.4), in order to investigate the effect of using a trigger model on individual turn pairs, the program previously used was applied to compare the probabilities given to individual dialogue turns according to a simple trigram model and according to a trigger model alone. The turns showing the greatest ratio of cache model probability to trigram model probability were output, and their content (together with the content of the immediately preceding turn) studied. Some examples of such turn pairs are given in the table below, where the number preceding the text is the logarithm to base ten of the ratio of trigger model probability to trigram model probability for the second turn of the pair. The text in square brackets is the preceding dialogue turn, whilst the turn of current interest follows. The + and - signs indicate that the word preceding it had probability at least 10% greater, or less, respectively, according to the trigger model relative to the trigram model.

```
72.158 [ NO I CAN'T STAND TO WATCH IT I'M TRYING NOT TO
REALLY DRINK ] CHI+ CHI+ CHI+ CHI+ CHI+ CHI+ CHI+ CHI+
CHI+ CHI+ CHI+ CHI+ CHI+ CHI+
  50.813 [ NO VICKI ] BOING+ BOING+ BOING+ BOING+ BOING+
BOING+ BOING+ BOING+ BOING+ BOING+ BOING+ BOING+ BOING+
  48.809 [ PUT THEM ON THERE THANK YOU ] DOR+ DOR+ DOR+
DOR+ DOR+ DOR+ DOR+ DOR+ DOR+ DOR+ DAD-
  33.587 [ DIDN'T SHE ] TICKLE+ ICKLE+ ICKLE+ ICKLE+
ICKLE+ ICKLE+ ICK+
  31.539 [ SHALL WE DO OUR SECTION WORK AS WELL ] CHING+
CHING+ CHING+ CHING+ CHING+ CHING+ CHING+
  27.680 [ SEEM LIKE ] GEHT+ DIR+ WOHIN+ KOMMEN+ SIE+
26.662 [ HERE WE GO ] HOUSE- HOME- BABY- NAPPY+ NAUGHTY
COT+ BED- BATH- RATTLE+ CRIB+ BROTHER- SISTER- AUNTY+
UNCLE+ MUM- DAD- GRANDPA+ STORIES+ BOYS- GIRLS- HATS+
COATS+ SHOES- GLOVES+ SMILES+ FROWNS+ TEARS+ JOY+ MEALS+
WHEELS+ CARDS- GAMES+ CHRISTMAS- EASTER+ CHURCH- LOVE-
HOLIDAYS+ CHORES+ OUTINGS+ CLEANING- CARPET TABLES SOFAS+
CHAIRS BEDTIME+ STORIES+ PRAYERS+ BOOKS- BAKING+ COOKING-
PASTRY+ CAKES+ BREAD- SWEETS+ FIREWORKS+ CRACKERS+
BIRTHDAY- TREATS+ PARTIES+ THEATRE+ FLICKS+ PANTO+
TRICKS+ MAGIC+ SHOWS PUNCH+ AND- JUDY- SWIM+ DANCE- SKIP+
  26.467 [ YOU NEEDN'T SING ON IT SHUT UP ] TWEET+ TWEET+
TWEET+ TWEET+ TWEET+
```

```
  25.369 [ COME ON I DON'T KNOW ] DUD+ DI+ DUD+ DI+
DUDDILY+ DID- DIDDLY+ DUD+ DI+ DUD+ DI+
  25.118 [ I LOVED HIM CARNALLY OH IT'S SOMETHING ]
CARNALLY+ SEMI+ CARNALLY+ CYRIL+ CONNOLLY+ NO- SEMI+
CARNALLY+ CYRIL+ CONNOLLY+
  22.854 [ THE THE QUESTION IS THAT HONOURABLE MEMBER TO
BRING IN HIS BILL SAY AYE AYE THEY AYES HAVE IT THE AYES
HAVE IT WHO WILL CONFIRM BRING IN THE BILL ] MR- ROBERT+
MCLELLAN+ MRS- MARGARET- EWING+ DOCTOR- NORMAN+ GODMAN+
MR- TAFFORD+ WIGGLY+ MR- RICHARD- SHEPPARD+ MR- DAVID-
CRIMBLE+ ALICE+ MAHON+ MR- DAVID- ALTON+ MR- BILL-
MICKEY+ AND- MYSELF-
  21.414 [ SULPHUR SULPHUR HYDROXIDE ] HYDROCHLORIC+
HYDROCHLORIC+ ACID+ HYDROCHLORIC+ HYDROCHLORIC+ ACID+
OKAY-
  21.358 [ HE WROTE PYGMALION AND SANG DOCTOR DOOLITTLE ]
ELIZA+ DOOLITTLE+ ELIZA+ DOOLITTLE+
  21.334 [ PARLEZ VOUS FRANCAIS ] AH- OUI+ UN+ PETIT+ UN+
PETIT+ POIS+
  20.376 [ OUI ] OUI+ MAIS+ C'EST+ FORMIDABLE+
  19.826 [ BRUCE ] HUP+ HUP+ HUP+ HUP+
  19.580 [ DO THIS YEAH NO WAY DID YOU GET IT ] MARCUS+
PEWTALL+ WOOD- PEWTALL+ MARCUS'S+ NICK- NAME- IS-
PEWTALL+ WOOD- PEWTALL+
  19.276 [ NO THIS WOMAN CAME ROUND LAST NIGHT ] ICH+
DEUTSCH+ SPRECHAN+ OU+ A- FRANCAIS+
  19.175 [ NAME BRITAIN'S LARGEST KNOWN MAMMAL ] LOCH+
NESS+ MONSTER+ MOOSE+ LOCK- NESS+ MONSTER+ RED- DEER+
  19.003 [ I BEG YOUR PARDON SOUNDS NASTY JA JA HABEN SIE
KINDER NEIN ] WASCHER+ WASCHER+ NUMMER+ KINDER+
```

Although it is possible that, in some of the above examples, one very unusual word triggers another unusual one (e.g. in the cases where foreign words – notably chunks of French or German – not adopted into "standard" English appear), a more likely explanation is that such words are so unusual in the BNC that their probabilities are being boosted within the trigger model up to the default 1/W value since no triggers are active at that point. However, as noted in section 6.4 above, some pairs of items of mathematical and chemical terminology **did** appear in the lists of the "best" trigger pairs found in the training data. Thus, cases like the example where "Sulphur hydroxide" (sic) is followed by "Hydrochloric acid" may be genuine situations where active trigger pairs are enhancing a word's probability (and hence the probability of the dialogue turn) compared with that given by the simple trigram model. However, caution should be exercised before concluding that any pair of turns showing a semantically (or otherwise) plausible pair of words which might be expected to form a

pair of triggers. It should be remembered that the trigger-based model works purely on statistical associations found in the training data. A human eye may spot associations between words, based on wider or specialised knowledge, which we could not hope a model based on a strictly limited number of trigger pairs and exposed to a severely limited quantity of training material to acquire. For example, consider the following turn pairs, all relatively highly-ranked for probability enhancement by the trigger model (with the initial number being the logarithm of the ratio of the second turn's probabilities with respect to the trigger-based model to that given by the simple trigram model) :

13.992 [JUST ANY METALS] OH- IRON+ ALUMINIUM+ BRASS+ COPPER+ STEEL+ MAGNESIUM+ SODIUM+ POTASSIUM+ CALCIUM+

11.503 [HURRICANE] TYPHOON+ HURRICANE+ WAVES+

11.427 [NO I DON'T MEAN JUDY GARLAND I MEAN] LIZA+ MINELLI+ (Liza Minelli is Judy Garland's daughter. However, would these names appear in the list of triggers selected ?)

In the first of these, it might be expected that the word "metals" would trigger one of the names of specific metals in the following turn. Similarly, in the second example, "hurricane" might have been expected to trigger itself and/or the other words of the second turn associated with extremely severe weather. However, none of the revelant trigger pairs occur within the lists (based on those trigger pairs showing best mutual information with respect to the training dataset) used for the appropriate experiment. The third example is one which illustrates that we should not assume that the model has any "higher level knowledge" which a person may possess. Although a human familiar with American cinema may know that Liza Minelli is Judy Garland's daughter (and hence, from a psycholinguistic perspective, for such a person , "Judy Garland" may act as a "prime" or trigger for "Liza Minelli") there is no reason to assume that, based on limited statistical evidence, a model incorporating only those trigger pairs which are believed to be the most generally useful would acquire any such association. Each of these names may occur only a few times within the training corpus and indeed, none of the possible pairings between the two of them occurs in the list of triggers used for that experiment. Similarly, in the case :

10.990 [IF JUST TAKE JUST THINK OF SINGERS AT THE MOMENT SOPRANO IS THE HIGHEST THEN AN ALTO] SOPRANO+ THEN- AN- ALTO+ TENOR+ BASS+

only a person (or machine ?) with some knowledge of, or familiarity with, musical terminology would make the association between "singers", "soprano" and "alto" with "tenor" and "bass". All of these are relatively rare words and it is not very surprising that none of these appear in the set of trigger used in that experiment.

Some other such examples requiring "higher level knowledge" (e.g. a turn pair featuring "Pygmalion" and "Eliza Doolittle", and cases where one or more words in French or German in the first turn are followed by more in the second of the pair) appear in the earlier list of turns most highly ranked in this manner.

In fact, it would appear that only a relatively small proportion of those turns most highly ranked (by improvement in probability through use of the trigger-based model) contain a target word which both occurs in the set of triggers used for that particular experiment and for which a "semantically salient" triggering word appears in the preceding turn. Some such examples are :

11.556 [AND SAUSAGES AND HORS D'OEUVRES] SAUSAGES HORS+ D'OEUVRES+ ("Sausages" acts as a self-trigger here)

7.447 [PECUNIARY] NON+ PECUNIARY+ (Self-triggering of "pecuniary")

7.447 [NON PECUNIARY] NON+ PECUNIARY+ (As above)

Here we have an example of an extremely rare word, but where it does occur within the training data, it normally also occurs in either the preceding or following turn (it was found 10 times in consecutive turns in the training data and there were 415190 triggering events where it was found in neither turn, but it only appeared 3 times in isolation). Hence the pair where "pecuniary" triggers itself has a reasonably high mutual information and is therefore included in the set of triggers pairs incorporated into the model.

In many of the other cases of turns highly ranked by this method, we are again seeing the probabilities of very rare words being enhanced unrealistically due to them being given the default probability when no triggers are active in the framework of the trigger based model. This is a possible weakness of the trigger model as it is formulated at present.

6.6 Summary

The results of the experiments described in this chapter show that incorporating a contribution from a trigger-based model can be of benefit to a statistical language model for both dialogue and text data from the BNC. However, particularly in the case of dialogue data, the results are rather disappointing compared with those obtained for cache-based models in chapter 5. The interpolation weights for the trigger-based component are very close to zero, suggesting that the interpolated models are mainly relying on the simple trigram component and gaining little information from the trigger model. Although the reasons for these observations are not yet fully understood, one possibility is that the influence of correlations between distinct words (exploited by trigger models) is strongest at very short ranges (of the order of 10 to 20 words), whereas the influence of repetitions of the same words (i.e. re-occurrences) is strongest at slightly longer ranges (of the order of 100 words). This is broadly in agreement with Rosenfeld's findings on "long-distance bigrams" and "distance-based triggers" for text data from the Wall Street Journal corpus (Rosenfeld 1996). Another possible reason for the relatively poor performance of the trigger-based models in this study is that they have not incorporated enough trigger pairs – for example, Rosenfeld & Huang (1992) used 620 000 triggers when modelling text from the Wall Street Journal. When their trigger-based model was interpolated with a simple trigram model, a 10% reduction in perplexity over a baseline of the trigram model alone was obtained. However, as noted by Rosenfeld (1994, 2000a), the computational demands of trigger models greatly increase as the number of triggers being employed is increased, and the models incorporating a few thousand triggers described in this present study seem to be approaching the limit of complexity which can be dealt with in a reasonable amount of computational time on the hardware available locally. Particularly when only a very short window, such as the most recent

dialogue turn, can be referred to, it is unlikely that many triggers will be active at any given instant, reducing the utility of the trigger model. When no triggers are active at all, the trigger model gives all possible words the same default probability which in many cases is highly unrealistic.

Furthermore, Rosenfeld (1996) notes that combining trigger and trigram models by linear interpolation is necessarily sub-optimal. Indeed, the results of such a model trained on almost 38 million words and tested on 70 thousand words of text from the Wall Street Journal corpus, incorporating 620 000 triggers, yielding only a 10% improvement in perplexity over the baseline of a simple trigram model (Rosenfeld & Huang 1992, Rosenfeld 1996).

It had been hoped that, in analogy with the psycholinguistic phenomenon known as priming, the trigger pairs found to be "useful" in the statistical models (using the automated selection process involving mutual information) would also prove to be "interesting" linguistically. Although some interesting trigger pairs, both from a semantic/lexical perspective and from a structural point of view (e.g. one pronoun triggering a related one, such as "she triggering "her") for the dialogue material, many of the trigger pairs found were either between extremely common words, or between rather rare words – the latter cases believed to be consequences of the nature of parts of the training dataset.

Conversely, many examples have been found where semantically associated pairs of words – clear to a human observer - have not genuinely benefited from the use of a statistically-trained trigger model. This is almost certainly due to the relevant pairs of related words not co-occurring (or being "co-absent") sufficiently frequently within the training data. In many cases, the human observer is making use of higher-level knowledge or intuition about language to which the trigger model – using a restricted number of trigger pairs and statistically-trained on a limited amount of data – has no access. Rosenfeld (2000b) has proposed an interactive strategy, "interactive feature induction", where human knowledge plays a complementary role with statistical, data-driven methods.

These observations suggest that, whilst trigger-based models may be of some benefit to the modelling of dialogue, their value may be limited on currently-available computational hardware unless the models can be incorporated into some type of hybrid model in a way superior to linear interpolation with a simple trigram model.

Furthermore, some of the results presented in this chapter illustrate that the dialogue material within the BNC, whilst quite extensive in size and diverse in its sources, does not appear to represent modern British English dialogue in a genuinely balanced manner – some very restricted topics seem to be disproportionately represented. This has implications for training statistical language models such as those based on word trigger pairs and the applications to which such models can be put with success.

Chapter 7 Experiments using Language Models Based on Clusters

7.1 Motivation & Overview

As discussed in chapter 2, methods based on clustering documents can be a useful means of allowing a statistical language model to adapt to suit the utterance or text currently being considered. In the context of a dialogue system, pairs of successive dialogue turns have an obvious importance : one turn of each pair will be the machine's own utterance, so will be known by it with certainty. In the light of perfect knowledge of the previous turn, a decision on which cluster model (or combination of models) would be most appropriate for application to the next turn (the user's turn) can be made. In this chapter, various strategies are proposed and tested for this purpose, and the results obtained compared. Also included is a so-called "oracle" strategy, which is a cheat – the decision regarding which cluster model should be applied to the second (the user's) turn of a pair is based on information about the second turn which, for a real dialogue system, could not be known in advance. However, the results from this strategy can be used to give a bound on the best performance which could possibly be hoped-for by this cluster-based approach if it were somehow possible to predict the correct cluster for the second turn of the pair with absolute certainty from complete knowledge of the corresponding first turn.

For each experiment, a trigram language model for second turns of pairs was constructed from the data in each cluster. This was then tested on data held back from the training set. A hybrid model, interpolated with the ordinary trigram model (which had been trained on the full training set, rather than on just the content of one cluster) was produced. The optimal interpolation weights for each cluster were computed using the Expectation-Maximisation algorithm (Dempster, Laird & Rubin 1977, Jelinkek 1990) with data reserved for this purpose, and an average set of interpolation parameters calculated and applied for all the cluster models in that particular experiment. Comparison was made between the perplexity scores for the individual cluster models, the "baseline" ordinary trigram model and the interpolated trigram-cluster model when applied to data held back for use in testing.

As a contrast, a strategy based on clustering whole dialogues and using the resulting clusters to produce "weighted mixture" models was also investigated.

7.2 Clustering Experiments on Ordinary BNC Dialogue Data – Using Whole Dialogues and a "Mixture of Clusters" Model

An experiment was carried out to investigate the effects of basing a language model on a weighted mixture of language models, each based on a single clusters of dialogues. Whole dialogues from the BNC were first put into 10 clusters, using the "k-means" approach described in section 2.2.3 and the lexically-motivated metric proposed by Robertson and Spärck Jones (1997). A trigram language model was then constructed for each such cluster. These were then interpolated with an ordinary trigram model trained on dialogue in two different ways. Firstly, optimal interpolation parameters were computed using 300 000 words of dialogue data reserved for this purpose, and the resulting "static mixture" model applied to test data. The other approach – an attempt to investigate how this type of clustering could enable the model to adapt to the material to which it was currently being applied – re-calculated the optimal interpolation parameters (or "weights") over each 10% of the test data, and the resulting "temporary mixture" model applied to the following 10% of the test data. I.e. if the 10 approximately equal-sized portions of the test data are labelled 0, 1, 2, ... , 9 respectively, then the interpolation parameters computed using data from portion 0 are applied to portion 1, those computed using portion 1 are applied to portion 2, etc. In each case, the relevant interpolation parameters were calculated using the *interp* program in the CMU Language Modelling toolkit, using the Expectation-Maximisation Algorithm (Dempster, Laird & Rubin 1977, Jelinek 1990). In the absence of a previous portion, the weights applied to portion 0 of the test data were those computed for the static mixture model. Such a "mixture model" approach has previously been used by Clarkson & Robinson (1997) and Iyer & Ostendorf (1999).

For comparison purposes, in a ten-fold cross validation experiment, the perplexity of an ordinary trigram model trained on approximately 7 million words of dialogue

material from the BNC (see section 4.6) ranged between 168.31 and 202.28, with the mean value being 185.98.

If a single "static mixture" model (trained on one set of reserved data) was applied to the entire set of test data, a perplexity of 185.77 was obtained.

The results for the 10-stage adaptation process ranged between 159.26 and 220.09, with a mean value of 186.50. This model does not contain any features which allow it to adapt to the nature (e.g. the topic) of the data at a very local level – modifiying the weightings just after every 10% of the data is rather crude - and so there is little reason to expect it to be much better than the ordinary trigram model.

These results suggest that relatively little extra value can be obtained from clustering whole dialogues and incorporating models based on these clusters into a mixture model – at least with respect to this BNC dialogue data. The approach of updating the interpolation weights in stages is probably best-suited to use in cases where a single long document (or long dialogue), where the theme or topic of conversation changes gradually, is being considered. There is no real reason to suppose that dialogues from distinct files within the BNC – particularly if the ordering of the files has been randomised - should be in any way connected, so this process of adapting the weights only 10 times over the entire test set would probably not be expected to yield great benefit. Adaptation of the weights with each new file or each new dialogue might be more appropriate.

As was noted in section 4.3, some of "dialogues" in the BNC are rather long, and many are not well-balanced in terms of both speakers uttering approximately equal proportions of the total number of words. Thus, a substantial part of the BNC "dialogue" material really consists of chunks of monologue, interspersed with short comments from another speaker. Such examples are probably not typical of the type of highly interactive dialogues of greatest interest to speech technologists. Thus, for the remainder of this chapter, we will concentrate on cluster-based models using the DRT dataset – pairs of successive relatively short dialogue turns from the BNC.

7.3 Clustering Experiments on DRT Dialogue Data Using a Lexically-Motivated Similarity Metric

Following the argument proposed earlier that, as far as "dialogue" data from the BNC is concerned, the most important correlations between successive dialogue turns and most notable structural (rather than topic-based) features are to be found between turns of relatively short lengths, we have again focused on the Dialogue Reduced Turn (DRT) data constructed from the dialogue material in the BNC as described in section 4.5. In each experiment, dialogue turns (rather than whole dialogues) were first allocated to clusters using the "k-means" approach with the "lexically-motivated" metric proposed by Robertson & Spärck Jones (1997), as described in section 2.2.3.1. Language models were produced for each of the resulting clusters.

Three sets of experiments have been carried out using different strategies for determining the clusters and, during the "recognition" phase, selecting which cluster language model is appropriate. Additionally, a set of "oracle" experiments, as described in section 7.1 above, was also carried out. Each of these strategies makes turns forming a consecutive pair.

Experiment type "F" : The clusters were built according to the content of both first and second turns of each turn pair in the training set, i.e. treating each turn pair as a single entity. Once all the training material had been assigned to clusters, the mean value of the metric for each cluster was computed along with a language model appropriate for that cluster. In testing, a model was chosen for the second turn of the pair according to which cluster the corresponding first turn would be assigned – i.e. the value of the metric was calculated for the first turn of the pair currently under consideration and the cluster with mean closest to that value found. The language model appropriate for that cluster used for the second turn of the current pair. This model was interpolated with the "full" language model for second turns. Once results had been compiled for each of the 10 clusters, an average perplexity, weighted according to the sizes of the clusters, was computed. This approach to modelling turn pairs using clusters effectively assumes that the first and second turns of any pair are alike and cluster in the same way, so that use of either turn is a good way of predicting the content of the other.

Experiment type "T" : Similar to type "F" above, but the clusters were built according to the content of the first turns of the pairs only. Each pair of turns was effectively split into its two constituent turns, the first turn being put into a particular cluster (say cluster number i) and the second turn being put into a special cluster constructed to "shadow" the cluster used for the corresponding first turn (in this case "shadow cluster number i). Thus, the turns ending up in shadow cluster i are the second turns of pairs of which the first turns are in main cluster i. A language model was constructed for each of the shadow clusters. In testing, the most appropriate cluster for the first turn of each pair was calculated and the corresponding shadow cluster model applied to the second turn of the pair. This effectively assumes that the content of the first turn of a pair is a good predictor of the content of the second turn, but not necessarily the other way round. It does not make any assumption about the actual content of the two turns being alike, only that one is a good predictor of the other.

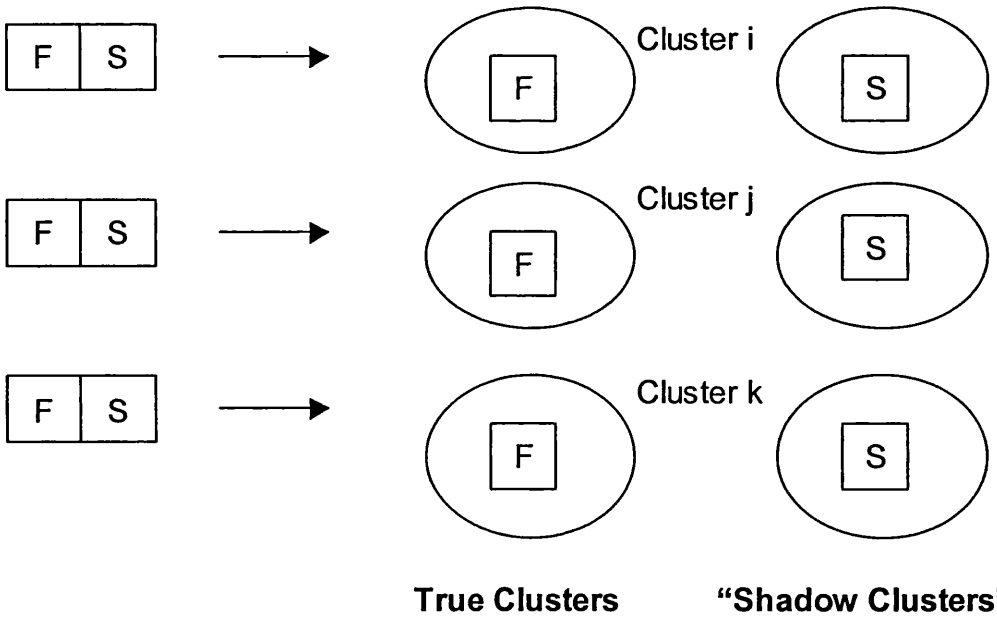


Figure 7.1 : Production of clusters and "shadow clusters" for first and second turns of pairs in a "T"-type experiment. The clustering algorithm is applied to first turns alone, but if such a first turn of a pair is allocated to cluster i, then the corresponding second turn of that pair is allocated to shadow cluster i. "F" and "S" refer to the first and second turns of a pair respectively.

Experiment type "R" : Similar to type "F" above, but only the content of second turns of the pairs was used in constructing the clusters. In an analogous manner to the "T" type experiments described above, but reversed, each second turn of each pair was allocated to a particular cluster (say cluster j) and the corresponding first turn of that pair allocated to the corresponding "shadow" cluster (shadow cluster j in this case). Thus, in this case, shadow cluster j contains the first turns of those pairs of which the second turns have been assigned to normal cluster j. The mean value of the metric was computed for each shadow cluster. During testing, the appropriate shadow cluster was found for each first turn of a pair. The language model for the corresponding normal cluster was then used for each second turn of that pair. The assumption that it is appropriate to cluster second turns and then produce "shadow clusters" for the corresponding first turns in effect assumes that, in a case where there was no direct knowledge of the content of the first turn of a pair, the second turn is a good predictor of the content of the unknown first turn. Although this may at first sound unrealistic, a system based on this principle could be used to retrospectively reconstruct or "repair" an unmonitored or corrupted (e.g. by noise) turn using knowledge of the turn which immediately follows it.

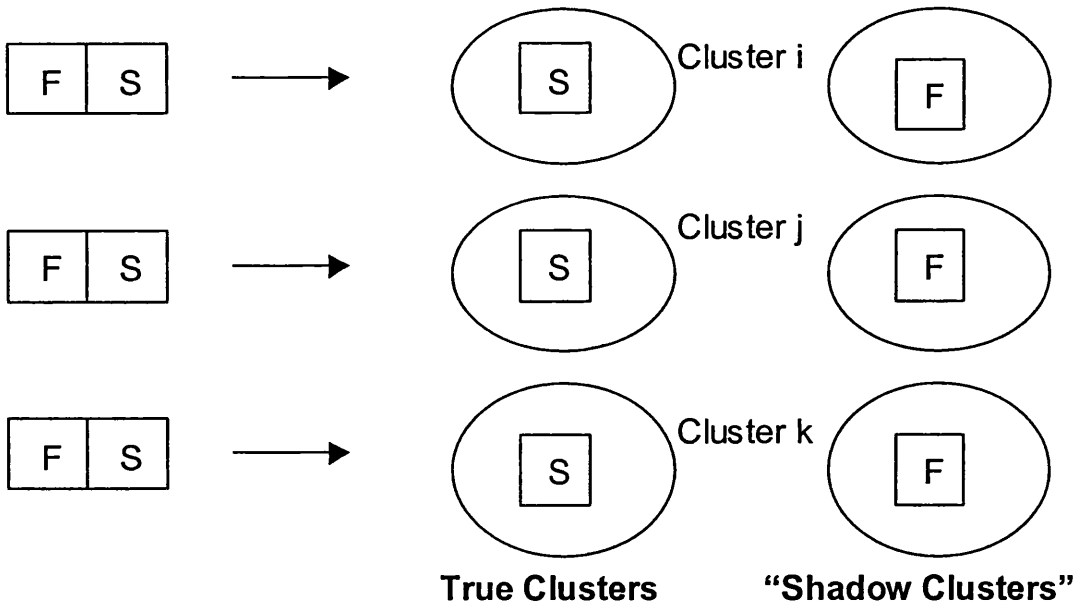


Figure 7.2 : Production of clusters and "shadow clusters" for first and second turns of pairs in an "R"-type experiment. The clustering algorithm is applied to second turns alone, but if the second turn of a pair is allocated to cluster i, then the corresponding first turn of that pair is allocated to shadow cluster i. "F" and "S" refer to the first and second turns of a pair respectively.

Experiment type "O" : The "oracle" experiments, where only the content of second turns of pairs was used to build the clusters and, in testing, the language model for the second turn was chosen according to which cluster that turn would be assigned - i.e. the metric was computed for the second turn of the pair currently under consideration and its value compared with the mean values the metric for the clusters produced from the training data. The language model constructed for the cluster with mean closest to the value of the metric for the current second turn was applied to that turn. This process relies on knowledge of the turn currently being tested and, in terms of any real dialogue system, is therefore a cheat. However, unlike the other three approaches, it does not make any implicit assumptions about correlations between the content of the two separate turns of each pair.

7.3.1 Dependence of Model Perplexity on Number of Clusters Used

For each type of experiment, an investigation was carried out on how the perplexity of the model depended on the number of clusters used, both for the cluster model alone, and for a hybrid model produced by interpolating the cluster model with the trigram model used as a "baseline" for comparison. In each case, the standard BNC dialogue lexicon of 50 000 words was used. In each experiment, perplexity and logprob scores were calculated for each cluster, then a weighted mean logprob calculated across the M clusters, taking account of the different numbers of words in the distinct clusters. From this, a "weighted average perplexity" score for that set of cluster models was calculated.

If cluster i contains n_i words and gave a perplexity score pp_i, corresponding to a logprob lp_i, then $lp_i = \log_2(pp_i)$. The weighted mean logprob across the M clusters is then :

$$WMLP = \frac{1}{n} \sum_{i=1}^{M} (n_i . lp_i)$$

where $n = \sum_{i=1}^{M} n_i$.

The weighted average perplexity is then :

$$WAPP = 2^{WMLP}.$$

The summaries of results from these experiments are shown in tables 7.1 to 7.4 below. In general, the models using a single cluster alone tended to have rather variable perplexities, often rather high values. This is largely due to the limited amount of data available for training each single cluster model. It can be seen that, for a given type of experiment, the larger the number of clusters used, the higher the average perplexity of the models using data from any one cluster alone. (In fact, the case of using the simple trigram model alone can be considered as using just one cluster, fitting-in with this trend.) This is because the training set used in such cases is made smaller – there is a fixed amount of training data in all, so the larger the number of clusters used, the smaller the average amount of data in each cluster. Generally speaking, the interpolated models had lower perplexities than those of the corresponding single cluster or simple trigram model, showing that the use of a cluster model could be a useful supplement to a standard trigram model. As expected, the number of clusters has very little effect on the perplexity of the simple trigram model – changing the clusters being considered only makes relatively small adjustments to the way the training and test data sets are partitioned. However, there is a small but noticeable improvement (i.e. decrease) in the average perplexity of the interpolated trigram-cluster models as the number of clusters is increased. This may be due to a finer classification of the data being possible when the number of clusters is larger. This trend is similar to that found by Clarkson & Robinson (1997) for models trained on mixed text and spoken material from the BNC.

Number of Clusters	Weighted Average Perplexity			Improvement
	Cluster-based model only	Simple trigram model only	Interpolated model	
5	246.695	218.539	213.325	2.39 %
10	265.991	218.542	212.994	2.54 %
20	307.745	218.544	212.683	2.68 %
40	335.103	218.542	211.908	3.04 %

Table 7.1 : Comparison of Weighted Average Perplexities of Cluster-Based, Simple Trigram and Interpolated Cluster-Trigram Language Models using different numbers of clusters. "F"-type experiment, full 50 000 dialogue lexicon used. Percentage improvement values quoted are for the interpolated models relative to the corresponding baseline simple trigram model. The small variation in the perplexity scores of the simple trigram models with the number of clusters used is due to the accumulation of rounding errors during the averaging process, which are not identical when the dataset is partitioned in different ways.

As was expected, the "oracle" experiment (the cheat) yielded lower perplexities than the others and, when interpolated with the simple trigram model, the largest improvements over the baseline. These figures represent the greatest reduction which could be hoped-for using this type of approach, i.e. a "bound" on the best possible performance using this type of model. This improvement – of up to about 13% of the perplexity of the baseline model in these experiments - is encouraging and suggests that cluster-based approaches may indeed be valuable in the statistical language modelling of dialogue. However, the "oracle" methodology relies on information within the second turn of the pair and is therefore invalid for predicting the content of that turn.

In terms of average perplexity scores, there was little difference between the results of the other types of cluster experiments.

Number of Clusters	Weighted Average Perplexity			Improvement
	Cluster-based model only	Simple trigram model only	Interpolated model	
5	262.157	218.542	213.625	2.25 %
10	287.521	218.544	212.919	2.58 %
20	324.136	218.541	212.637	2.70 %
40	359.095	218.541	212.155	2.92 %

Table 7.2 : Comparison of Weighted Average Perplexities of Cluster-Based, Simple Trigram and Interpolated Cluster-Trigram Language Models using different numbers of clusters. "T"-type experiment, full 50 000 dialogue lexicon used. Percentage improvement values quoted are for the interpolated models relative to the corresponding baseline simple trigram model.

Number of Clusters	Weighted Average Perplexity			Improvement
	Cluster-based model only	Simple trigram model only	Interpolated model	
5	258.916	218.543	213.448	2.33 %
10	297.410	218.701	213.215	2.51 %
20	323.220	218.542	212.273	2.87 %
40	360.287	218.543	212.075	2.96 %

Table 7.3 : Comparison of Weighted Average Perplexities of Cluster-Based, Simple Trigram and Interpolated Cluster-Trigram Language Models using different numbers of clusters. "R"-type experiment, full 50 000 dialogue lexicon used. Percentage improvement values quoted are for the interpolated models relative to the corresponding baseline simple trigram model.

Number of Clusters	Weighted Average Perplexity			Improvement
	Cluster-based model only	Simple trigram model only	Interpolated model	
5	222.337	214.631	206.930	3.59 %
10	242.789	214.624	198.882	7.33 %
20	255.405	214.618	191.290	10.87 %
40	274.136	214.505	188.148	12.29 %

Table 7.4 : Comparison of Weighted Average Perplexities of Cluster-Based, Simple Trigram and Interpolated Cluster-Trigram Language Models using different numbers of clusters. "O"-type (Oracle) experiment, full 50 000 dialogue lexicon used. Percentage improvement values quoted are for the interpolated models relative to the corresponding baseline simple trigram model.

7.3.2 Dependency of Language Model Perplexity on Size of Lexicon Used

Experiments were also carried out, using a fixed number (10) of clusters in each, to investigate the effect of varying the size of the lexicon used on the perplexity of the resulting cluster-based language models. The different lexica were only used in the construction of the clusters and did not affect the trigram model for the full dataset. The results are summarised in tables 7.5 and 7.6 below.

Size of Lexicon (words)	Weighted Average Perplexity			Improvement
	Cluster-based model only	Simple trigram model only	Interpolated model	
50 000	265.911	218.542	212.994	2.54 %
5 000	267.833	218.545	212.995	2.54 %
500	264.749	218.540	213.073	2.50 %
50	281.547	218.544	213.250	2.42 %

Table 7.5 : Comparison of Weighted Average Perplexities of Cluster-Based, Simple Trigram and Interpolated Cluster-Trigram Language Models using different sizes of lexicon. "F"-type experiment, 10 clusters used in each case. Percentage improvement values quoted are for the interpolated models relative to the corresponding baseline simple trigram model.

Size of Lexicon (words)	Weighted Average Perplexity			Improvement
	Cluster-based model only	Simple trigram model only	Interpolated model	
50 000	242.789	214.624	198.882	7.33 %
5 000	242.998	214.631	198.884	7.34 %
500	235.693	214.674	200.763	6.48 %
50	245.971	214.688	198.060	7.75 %

Table 7.6 : Comparison of Weighted Average Perplexities of Cluster-Based, Simple Trigram and Interpolated Cluster-Trigram Language Models using different sizes of lexicon. "O" (oracle)-type experiment, 10 clusters used in each case. Percentage improvement values quoted are for the interpolated models relative to the corresponding baseline simple trigram model.

As can be seen from the above, there appears to be little effect on model perplexity - with the exception of the perplexity of the uninterpolated cluster model in the "F" type experiment when the very small (50 word) lexicon was used, which was slightly higher – due to varying the lexicon size. When the content of individual clusters was investigated (see sections 7.6 and 7.7 below), in many cases the cluster for a given turn seems to have been determined by the presence or absence of one or more words from a small set of relatively common words. This may explain why such little variation of cluster model perplexity was found as the size of the lexicon used was varied.

7.4 Experiments Using an Entropy-Based Clustering Metric

In a similar manner to the experiments described in section 7.3 above, experiments were carried out to investigate the effect of clustering the dialogue turns of the DRT dataset using the entropy-based (or, equivalently, perplexity-based) metric described by Carter (1994a,b). The full 50000 word dialogue vocabulary was used as lexicon. Concentrating on "O" (Oracle) type experiments (see section 7.3 above), language models were constructed for each of M clusters (for M = 5, 10, 20 or 40). The perplexities of these models were then calculated with respect to appropriately-clustered data held back during the building of the models. The cluster models were also interpolated with a simple trigram model, the optimal interpolation parameters being calculated with respect to data reserved for this purpose. The results are summarised in tables 7.7, 7.8 and 7.9 below.

It can be seen that the weighted average perplexities of both the cluster models and the optimally-interpolated trigram-cluster models decrease as the number of clusters used in the experiment is increased. In all cases, there are big differences in the sizes of the individual clusters, with it being commonplace for the largest cluster in one particular experiment to be as large as all the other clusters put together. In most (but not all) cases, the largest perplexity for both the purely cluster-based and the interpolated trigram cluster models are found for the largest cluster of an experiment.

It can also be noted that for small numbers of clusters, the entropy-based metric performs better than the lexically-based one, whereas the converse is true if larger numbers of clusters are employed. This will be further discussed in section 7.5 below.

Number of Clusters	Weighted Average Perplexity (interpolation weight for cluster model)			Improvement
	Cluster-based model only	Simple trigram model only	Interpolated model	
5	222.006	214.478	202.652 (0.457035)	5.51 %
10	221.087	214.413	200.571 (0.469487)	6.46 %
20	218.591	214.638	199.280 (0.460405)	7.16 %
40	217.063	214.643	199.028 (0.464878)	7.27 %

Table 7.7 : Weighted average perplexities of cluster-based, simple trigram and interpolated trigram-cluster models for DRT dialogue data from the BNC, where the clustering was carried out using the entropy-based metric. "O" (Oracle) type experiments using the full 50000 word dialogue lexicon. Rotation 0 only.

Number of clusters	Size of largest cluster (words)	Size of smallest cluster (words)	Peplexity of cluster-based model alone		
			Weighted average	Highest value for one cluster	Lowest value for one cluster
5	381010	22497	222.006	263.64*	109.29#
10	395807	5180	221.087	258.06*	87.09
20	413450	2159	218.591	250.14*	59.42
40	432138	350	217.063	335.04#	34.74

Table 7.8 : Summary of cluster sizes and perplexities of cluster-based models (during evaluation of trained models), rotation 0 only. "O"-type experiments with the entropy-based clustering metric and the full 50000 word dialogue vocabulary.
* occurs for largest cluster, # occurs for smallest cluster.

Number of clusters	Perplexities of simple trigram models alone		Perplexities of interpolated trigram-cluster models	
	Highest value for one cluster	Lowest value for one cluster	Highest value for one cluster	Lowest value for one cluster
5	270.49	85.09	254.773*	81.33#
10	266.82	68.16	251.186*	61.96#
20	261.85	51.97	246.447*	44.47#
40	256.13*	43.93	242.259*	29.476

Table 7.9 : Summary of range of perplexities found across clusters during evaluation of cluster models : simple trigram models and interpolated trigram-cluster models. "O"-type experiments using the entropy-based clustering metric and the full 50000 word dialogue vocabulary.
* occurs for largest cluster, # occurs for smallest cluster.

It would appear that the entropy-based clustering method produces one very large "dustbin" cluster for turns which are relatively unpredictable (and hence, when modelled, show a high perplexity). The turns of the other clusters are then, in relative terms, more predictable, and the language models for those clusters are thus of lower perplexity. (However, due to the relative sizes of the clusters, the weighted average perplexity across all the clusters is not necessarily lower than in cases resulting from the use of the lexically-based metric.) This would suggest that modelling the individual clusters of turns, where the turns have been identified and constructed using the entropy-based metric, could be of significant value in the modelling of dialogue.

7.5 Comparison of Entropy-Based and Lexically-Based Clustering Methods

A set of experiments, using all 10 cross-validations across the data, was carried out to compare the effects of the entropy-based (Carter, 1994b) and lexically-based (Robertson & Sparck-Jones, 1997) methods of clustering dialogue turns. These were all performed within the framework of "O" (Oracle) type experiments.

Number of clusters	Lexically-based metric	Entropy-based metric
5	3.59%	5.51%
10	7.33%	6.46%
20	10.87%	7.16%
40	12.29%	7.27%

Table 7.10 : Average relative reduction in perplexity (with respect to baseline of simple trigram model) of interpolated trigram-cluster models using different numbers of clusters and the two different clustering metrics. "O"-type experiments.

Findings indicate that, when a small number of clusters is used, there is little difference in performance – in terms of the relative reduction in perplexity of an interpolated trigram-cluster model over that of the trigram model alone – between the lexically-based and entropy-based clustering metrics. When 5 clusters were used in each experiment, the entropy-based metric gives slightly larger reductions in perplexity than were obtained using the lexically-based metric. On the other hand, in

the experiments using larger number of clusters (see sections 7.3 and 7.4 above), greater benefit seemed to be achieved by increasing the number of clusters employed in the case of the lexically-based metric than when the entropy-based metric was used. This may be because, when the number of clusters allowed is large, the lexically-based metric allows finer discrimination between the content of individual clusters.

However, in terms of the computational time taken to allocate turns to the most appropriate clusters and subsequent construction of language models, the entropy-based approach is somewhat faster – at least for the present implementations on the available hardware. The experiments using the lexically-based metric with 40 clusters took approximately 24 hours per cross-validation rotation on the hardware available, whereas those using the entropy-based metric on average took only about 10 hours for each rotation. For the same type of metric and number of clusters, there was little difference between the time required for the different types of clustering experiments but, as might be expected, the CPU time used increases as the number of clusters used is increased – probably growing at a rate greater than linearly with respect to the number of clusters.

Number of clusters used	Time per rotation (in hours)			
	Experiment type			
	"F"	"T"	"R"	"O"
5	0.914	0.918	0.922	0.865
10	1.810	1.728	1.764	1.720
20	4.060	3.955	3.730	3.989
40	10.397	9.903	9.806	10.711

Table 7.11 : CPU time required per cross-validation rotation on a 2.0 GHz Pentium 4 PC with 512 Mbytes of RAM for cluster model experiments with the entropy-based clustering metric.

For the entropy-based approach, working within the framework of these "O"-type experiments, for each cross-validation rotation across the data, the clustering

produced one very large cluster – often containing more words than all the other clusters put together – which showed very high perplexity for each of the cluster, simple trigram and interpolated trigram-cluster models. Generally, the perplexities of all the other clusters were much lower for all 3 types of language model. This suggests that the method is producing one cluster which acts rather like a "dustbin". In this are placed turns which the language modelling framework cannot make reliable predictions – hence the high perplexity values. The content of the turns in each of the other clusters is more predictable, hence the lower perplexity values.

In the case of the experiments using the lexically-based metric, the clustering method did always produce one particularly large cluster – in some cases containing more words than all the other clusters put together – in a similar manner to the entropy-based metric. However, unlike the entropy-based case, the large cluster did not always give the highest perplexity scores

7.6 Qualitative Discussion on the Content of Individual Clusters

It would be hoped that the clustering process would produce clusters of turns which were in some sense coherent. Perhaps the individual clusters would contain turns which appeared to share a common theme, or at least contain words which were in some way related ? This would particularly be expected to be the case when the lexically-based metric is used for producing the clusters.

In the results of this study, the very wide range of cluster sizes made direct statistical comparison between the clusters impractical. However, in some cases, it was possible to made some qualitative observations about the content of individual clusters.

For the clusters produced using the lexically-based metric, there were several cases where a single word, or members of a small set of words, appeared in the majority of the dialogue turns within that cluster. For example, when 40 clusters were used, produced using the metric suggested by Robertson & Sparck Jones (1997), with 10 cross-validation rotations across the data, the following were observed :

Rotation 0, cluster 12 : "Mm" occurs very frequently.

Rotation 0, cluster 14 : The word "Aye" occurs very commonly – but not in every turn – along with "Yes", "Yeah" and other words of acknowledgement.

Rotation 0, cluster 15 : Most turns contain a word being spelt-out alphabetically (e.g. "G R O U N D") or a chemical formula being spelt out (e.g. Na Cl "N A C L", "Ammonium ... N H Four"), or else one of two presumed mis-transcriptions of "Dunno" ("Du N No" or "Du N")

Rotation 0, cluster 25 : Most turns contain the word "Ah"

Rotation 1, cluster 32 : Many turns contain an acronym (particularly ones containing a letter "F", e.g. RAF, FBI, MFI. HBF, ESF, VHF, NFU) or another word being spelt-out alphabetically (e.g. "F L O O D E D", "F R I A R", "M A N T U A", "V E R O N A"). Names of fighter aircraft (e.g. "Eurofighter", "Jaguar" and "Phantom") are also quite common.

Rotation 2, cluster 1 : "Yep" occurs in most turns.

Rotation 2, cluster 15 : "Thank" occurs in most turns.

Rotation 2, cluster 20 : "Alright" occurs in most turns.

Rotation 2, cluster 23 : "Sure" occurs in most turns.

Rotation 2, cluster 30 : "Those" occurs in most turns. Wh-question words (e.g. "Which", "When", "How") are also quite common.

Such trends are less obvious within the clusters produced using the entropy-based metric. This is perhaps to be expected, since the metric is not explicitly based on the lexical content of turns – although it is certainly not correct to say that the entropy-based metric takes no account of the words present : rather, the dependence on the lexical content is more subtle, via probabilities given by the language model

constructed for each cluster. However, some clusters did show certain trends in terms of the words contained in their turns, e.g. (when 40 clusters were used in each of 10 cross-validation rotations across the data) :

Rotation 1, cluster 2 : Very common words include "Do" and "Does", "Did" and "Didn't", "Can" and "Can't", "Could" and "Thank". There are also quite a high proportion of "Wh-question" words : "What", "Where", "How", etc., and a relatively large number of pronouns, especially "You".

Rotation 1, cluster 4 : Pronouns (including compound pronoun-verb forms) are particularly common, including "It", "Its" and "It's", "You", "You've" and "You're", "They", "They're" and "They've".

Rotation 1, cluster 5 : Very small cluster consisting of single word turns – either "Okay", "Right" or "Thanks"

Rotation 1, cluster 14 : All very short turns, mostly including "This", "That", "That's", "What", "What's", "Which", "Where" or "Where's". "Agreed" is also quite common.

So we see that clustering using the entropy-based metric can produce a similar result – in effect, clustering some turns with similar lexical content together – to the approach using the metric which is explicitly lexically-based.

7.7 What types of turn benefit most from cluster-based modelling ?

In a similar manner to the method used for the cache and trigger models (see sections 5.4 and 6.5) on individual turn pairs, a program was used to compare the probabilities given to individual dialogue turns according to a simple trigram model and according to the cluster-based model alone. The turns showing the greatest ratio of cluster-based model probability to trigram model probability were output, and their content studied.

The number preceding each turn is the logarithm to base ten of the ratio of probabilities, whilst the + and – signs indicate that the probability of the word preceding it has been increased (or, respectively, diminished) by at least 10% by using the cluster-based model relative to the probability given by the trigram model.

In some cases, with the lexically-based metric, it was possible to spot some common feature of turns which were rated much higher by the cluster-based model than by the ordinary trigram model, e.g.

"O" type clustering experiment using 10 clusters:

Rotation 9, cluster 0 :

```
  74.987 YEAH- WEAR- THEM+ DA+ DA+ DA+ DA+ DA+ DA+
  61.807 WE+ WE+ WERE+ SAYING- THE+ OTHER+ DAY- YEAH- WE
RECKON+ WE+ RECKON+ YEAH- SHE- LOOKS+ LIKE- OI+ OI+ OI+
YOU+ KNOW+ OI+ OI+ OI+ YOU+ KNOW+ THE+ FRAGGLES-
  52.624 DOWN+ I- DU+ N+ NO+ MY+ FRIEND+ TOLD- ME+ YEAH-
HE+ SAYS+ IT'S+ TWENTY+ POUNDS+ FOR+ A+ SERVICE+
  43.188 YEAH- BUT+ YOU+ KNOW+ YEAH- YEAH+ BUT+ THEN+
CLAIRE+ CAME+ HOME+ WITH+ ME+ COS+ WE+ MISSED+
CORONATION+ STREET+ DIDN'T- WE+ LAST+
  41.697 NO+ YEAH- YEAH+ SAD- REALLY+ AND+ AMANDA+
DIDN'T- LIKE+ OUR+ OWN+ SCHOOL+ COS+ SHE+ FINISHES+ HER-
TEACHER+ TRAINING+ COLLEGE+ THIS+
  32.348 POLO+ NECK+ JUMPER+ I+ WANT+ SOME+
  27.792 THAT+ PROGRAMME+ IT+ CAME+ FROM+ RED+ DWARF+ OR
WHATEVER+
  25.473 YEAH- OH+ HE'S+ VERY+ BUSY+ YEAH- LOTS+ OF+
PEOPLE+ THAT+ ARE+ FINDING+ IT+ VERY+ DIFFICULT+ AREN'T-
  25.223 YEAH- YOUR- FAULT+ HALF PAST+ THREE+ I- COULD-
OF+ GONE+ OUT+ EVEN- MORE+ THEN+ OH+
  23.748 YEAH- I'VE- GOT G+ C+ S+ E+ IN+
  23.224 YEAH- YEAH+ COS- I+ WAS+ STAYING+ IN+ HER- ROOM+
AND+ I+ WAS+ ILL+ AND+ MM+ YEAH- YOU- CAN+ ALWAYS+ BLAME-
IT+ ON+
```

"Yeah" occurs in most of the above sentences.

Rotation 9, cluster 2 :

```
  38.109 I+ KNOW- SO+ THAT+ WAS- THE+ PHONE+ GOING-
TWICE+ NOW+ IT'S+ DISGUSTING+ OH- I'VE+ GOT+ TA+ GET+
WHAT'S- IT+ NAME- NEXT- WEEK+ NEXT+ MONTH+ I- CAN'T
REMEMBER+ WHAT+ IT'S+ CALLED+ IT'S+ GOT ERICA-
HASSLEHOFF- I- MEAN+ ERICA- TUT- IN+ IT- ANYTHING+ OF-
INTEREST+ IN+ IT+ OH- OH- OH THERE+ WE+ GO+ OH- YES+ OH-
```

35.516 OH- GOD+ YEAH THERE'S+ ORANGE- JUICE+ BUT OH- YEAH- AN- ORANGE+ MM+ WHAT+ IS THIS+ OH- I SEE+ PEOPLE+ WITH+ FAMOUS+ NAMES- THERE+ IS+ A+ PERSON+ CALLED+ DAWN+ FRENCH+ JENNIFER+ SAUNDERS- JULIA+ ROBERTS- MANDY+ SMITH+ HEY-

35.194 HIM+ AND+ ANDY+ GOT- PISSED+ RIGHT AND+ HE+ COULD+ SAY+ OH- THERE'S+ A- FIRE+ ANDY+ THERE'S+ A+ FIRE+ ANDY+ AND+ GOT- BLOWN+ AWAY+ RIGHT+ ANDY'LL- SAY+ OH- OH- OH

31.608 MUMMY+ GOES- RYAN+ SHOULD HAVE+ BEEN+ SHOULD HAVE+ BEEN+ A- ACTOR+ RYAN+ RYAN'S- SICK+ ALREADY- LAST- NIGHT- THOUGH+ THERE+ WAS+ RYAN+ RYAN+ SAT+ THERE+ OH- MUMMY+ MY+ BELLY+ OH- OH- MUMMY+ I'M- GON+ NA BE+

31.309 LATEST+ OH- OH- SHE COMES+ DA- NA+ NA+ NA+ NA+ NA+ OH- OH- HERE- SHE- COMES+ OH- YEAH+ OH- OH- HERE- SHE- COMES+ HA- HA MUM+ WIND- IT+

29.533 BRYANT+ MAYBE- NO+ I DIDN'T- PUT+ PAUL- COULDN'T+ HAVE- BEEN+ HIM+ SO+ I'M+ GON+ NA SHOW+ DAD+ THAT- LIST+ THIS- WEEKEND+ SEE- IF+ HE+ SAYS+ OH- OH

Observe how common the word "Oh" is in the above. Proper names (particularly of people) and references to food are also common in this cluster.

Rotation 9, cluster 3

56.872 ALL- RIGHT- THEN YEAH- LIKE+ I+ WAS+ SAYING+ I+ GOT INTO+ A- MASSIVE+ ARGUMENT+ WITH+ HIM+ LAST+ NIGHT+ I+ HAD+ AN+ ARGUMENT+ WITH+ OSMAN-

48.666 ALRIGHT+ LISTEN+ RIGHT- LISTEN- LISTEN+ LISTEN+ LISTEN+ LISTEN+ RIGHT- I'M+ JUST+ THINKING+ OF IT AGAIN- RIGHT- YOU'VE+ PROBABLY+ HEAR+ ALL+ THESE- RIGHT- THERE'S THIS- THERE'S- THIS- MAN+ AND HE+ NO- I+ DON'T+ WAN+ NA SAY- THAT- ONE- YEAH+ THERE'S+ THESE+ THREE+ MEN+ AND+ THEY'RE- WALKING+ THROUGH+ THE+

40.782 BECAUSE+ THIS+ IS+ WHAT+ EMMA+ DID+ RIGHT- SHE+ SAID+ WHEN+ WHEN- MRS+ SAID+ WHY+ DID+ YOU+ ASK+ EMMA+ SHE GOES- COS+ WE ASKED+ YOU+ TO+ GO THE- CINEMA+ ALL- THE+ TIME-

39.861 KATHY+ IS- THE ONE+ YEAH- NO+ THAT'S- RIGHT- YEAH+ SHE+ IS+ KATHY+ I+ DU+ N+ NO+ I+ JUST+ SORT- OF+ GET+ THOSE+ TWO+ MUDDLED+ UP+ I DON'T+ KNOW+ WHY+ ANYWAY- KATHY WAS+ JUST+ TALKING+ RIGHT- AND

39.558 FEMALE+ MALE+ MALE+ REGIONAL+ ACCENT+ HAVEN'T+ A- CLUE+

38.827 ALL- RIGHT- LOOK- WHAT'S+ THE WHAT'S+ THE+ WHAT+ CAN- YOU+ PUT- IN+ YOUR LEFT- HAND+ BUT+ NOT IN+ YOUR+ RIGHT- YOUR+ RIGHT- ELBOW+ I+ MEAN- COME- ON- IT'S+ A- BIT- OBVIOUS+ WELL+ WHAT- CAN+ YOU+ PUT- IN+ YOUR LEFT- HAND+ BUT+ NOT IN+ YOUR+ RIGHT- YOU+ CAN+ PUT+ YOU+ CAN-

PUT+ YOU+ CAN- PUT+ BUT+ YOU- CAN'T+ PUT+ YOUR- LEFT-
HAND+ IN- YOUR+ RIGHT- ELBOW+ SEE- SHANE+ I+ THINK+ YOU+
SHOULD+ STICK+ TO+ THE+ KNOCK+ KNOCK+
38.820 THAT'S- RIGHT- BECAUSE+ WHEN+ WE- CAME+ HOME+ WE
CAME+ HOME+ ON+ A+ FRIDAY+ NIGHT+ I+ SAY+ AND- OH- THE+

"Right" and variants such as "Allright" are extremely common in the above.

Rotation 9, cluster 4 :

108.154 WHAT+ WERE+ YOU+ WHAT- WERE+ YOU+ HOW+ COME+ YOU+
WERE+ WAITING+ FOR- JIM+ TONIGHT- OUTSIDE+ AH+ COS+ I-
WAS- JUST- LUSTING- AFTER+ NO- NO- NO- NO- NO- NO- NO-
NO- YOU+ THOUGHT+ I'D+ GOT- OFF+ WITH+ HIM+ FOR+ FUCK'S+
SAKE+ OH+ DID+ LIZZIE- TELL YOU MM+ NO- NO- I I+ DON'T
KNOW+ I+ SORT+ OF+ THOUGHT+ SOMETHING MIGHT'VE+ HAPPENED-
OH+ YEAH+ WELL- YOU+ KNOW+ OH- YEAH+ NO- WELL WE+ WHAT-
YOU- WHAT+ WERE+ YOU+ NO- WELL I WAS+ I+ WAS+ JUST-
TELLING+ HIM+ SOMETHING+ AND+ I+ SAID+ I'M+ NOT+ TELLING+
YOU+ TILL- TEN+ THIRTY+ JUST- TO+ KEEP+ HIM+ IN+
SUSPENSE- I CAN'T- BELIEVE+ I KNOW+ HE'S+ SUCH+ A+ DICK+
WHY+ DIDN'T+ HE+ GO- AND- BUST+ THEM- MAYBE+ HE+ WAS-
BORED+ WITH+ BUSTING+ PEOPLE+ WHAT+
68.557 NO- NO- NO- NO- NO- NO- NO- NO- YOU'VE+ GOT+
CHILLI- ON+ YOUR- FINGER+ HEY+ ROB- DO YOU+ WAN+ NA+
LEND+ ME+ FIFTY+ P+ I'LL- WAIT- TILL+ HE+ GETS+ BACK+
NOW+ YOU'RE+ ALL+ WITNESSES+ INCIDENTALLY+ I+ OWE+ THIS-
MAN+ A TICKET+ FOR+ THE+ CONCERT+ I'LL- GIVE+ IT- TO+
HIM+
42.002 NO- INSY+ WINSY+ SPIDER+ CLIMBING+ UP+ THE+
SPOUT+
37.362 DO+ YOU+ RECKON+ THIS+ SNOW+ WILL- HOLD+ OUT+
UNTIL+ CHRISTMAS+ NO- NO- NO- IT WON'T HOLD- OUT+ UNTIL+
CHRISTMAS+ NORMALLY- YOU+ GET+ TWO+ OR+ THREE+ WEEKS+ OF+
A+ BAD- COLD+
34.845 NO- IT'S- OVER+ THERE+ SORRY+ OI- OI+ OI+ OI+
CAN- I THAT- THAT+ YELLOW+ PLEASE- NO- SORRY YELLOW+
JANE- CAN- I BORROW+ THE+ YELLOW- NO- JANE+ I- HATE-
DRAWING- PEOPLE+
33.509 NO- NO- NO- NO- NOTHING+ LIKE- THAT+ WHAT+
PISSED+ ME- OFF+ IS+ IS+ HE'S- HANGING+ ABOUT+ WITH+
JAMES+ AND+ THAT LOT+ NOW+ RIGHT+ BUT+ WHEN+ YOU+ TALK+
ABOUT+ JAMES+ AND+ THAT LOT+ TO+ HIM+ ITS- OH- THEY'RE+
A- BUNCH+ OF+

"No" is particularly common in the above turns, as are "Don't", "Didn't", "Couldn't", etc. Pronouns are also very common, especially "I".

Rotation 9, cluster 5 :

59.064 WHO+ ARE SITTING+ WAITING- FOR- POSTS+ IT'S- MOST+ LIKELY+ BE+ ONE- OF THOSE THAT+ WILL+ GET+ IT- BUT+ HE'S- CERTAINLY+ HE'S- MADE+ ENQUIRIES+ THERE'S- A+ LOT+ OF CHANGES+ IN+ AFOOT WE'RE- I- SHALL- DISCUSS+ IT+ LET- YOU+ KNOW+ ALL+ ABOUT+ IT+ ONE- ONE- DAY- IF+ ANYTHING+ COMES+ OF- IT+ BUT+ UNLESS+ YOU+ DO+ SOMETHING+ YOURSELF+ THERE'S- NO+ PROMOTION+

50.965 THREE- KINDS- OF+ INDUSTRY+ JOY- LISTEN JOY YOU'VE- GOT- YOURSELF+ A+ PAGE- FULL OF LINES+ I'M+ SICK+ AND- TIRED+ OF+ THESE- PEOPLE+ SHOUTING- OUT+ CHOOSE- A+ PROPER+ ONE- AND HAND- IT+ IN+ BY+ ONE- FIFTY+ TODAY- SO- YOU+ CAN DO+ IT+ IN+ YOUR+ LUNCH+ HOUR+ THREE- TYPES- OF+ INDUSTRY+

46.976 I+ WILL+ HAVE- ONE- OF- THEM+ KRAYS- YOU+ KNOW+ THAT- YOUNG+ LAD+ WHOSE- AYE+ HE'S+ MADE+ ONE- OF+ THE+ KRAYS+ HIS+ DAD+ HAVE YOU+ READ+ IT+ WELL+ HIS+ STEPDAD- HE'S+ HE'S+ FROM OUR+ VILLAGE+ HIM+ HE+ WRITES+ TO+ HIM+ AND+ HE+ GOES+ AND+ VISITS+ HIM+ AYE+ ONE- OF+ THE+

46.913 WHAT+ HAPPENED- TO- THEM+ TOO- EXPENSIVE+ BENSONS+ HAVE- GONE+ UP- AFTER+ ALL+ THEIR- ADVERTS+ ABOUT+ STAYING+ AT+ ONE- NINETY- NINE+ OR+ SOMETHING+ THEY'VE+ GONE+ UP- NO+ NO+ THEY'VE- GONE+ UP- AGAIN+ I+ CAN'T- BELIEVE+ I- ACTUALLY+ BOUGHT+ BENSON+ BUT- THEY- NEVER+ SAID+ THAT+ THEY+ DID+ THEY+ HAD- ALL+ THESE+ ADVERTS+ STILL- AT- ONE- NINETY- NINE+ THAT'S+ WORSE+ THAT'S- WORSE+ COS- ROLL- UP+ IS- A- MAN'S+ A- MAN'S+ CIGARETTE- BENSONS+ WERE NEVER+ ONE- NINETY- NINE+ NO+ I+ MEAN+ I+ MEAN LAMBERTS- OH+ RIGHT+ THEY'RE- ONE- NINETY- FIVE+ IN+ MY+

43.477 TO+ THE- MOVIES+ ACTUALLY- I+ HAVEN'T- RECENTLY ONE- GUY+ IN+ THE CHOIR+ SAID+ THAT- THE+ TINA+ TURNER+ ONE- IS- PROBABLY- ON- IN+ STRAWBERRY- HILL- AND+ HE+ LIVES+ NEAR+ THERE+ SO- YOU+ COULD- GO+ TO+ THAT- I'M+ QUITE+ INTERESTED+ IN- SEEING+

43.033 LIKE- THAT- AND+ THE+ THINK- THEY'RE+ BEING+ CLEVER- WELL+ LET+ THEM+ GET+ ON+ WITH+ IT+ BUT+ IF+ THERE- WEREN'T+ ANY+ MONEY+ FOR- THEM+ ONE- PARENT FAMILIES+ THERE SHOULDN'T+ BE+ ANY+ ONE- PARENT-

42.970 YEAH+ MY- DAD+ AND+ MY+ COUSIN+ THEY+ WERE- GON- NA+ MY- DAD+ SAID+ YEAH+ WELL- MY- ONE- NO+ MY+ DAD'S+ ONE- WAS CALLED+ RHINO- AND+ THE- OTHER- ONE- WAS CALLED+ ELEPHANT+ AND+ THEIR+ ONE- DIED+ THEY+ KEPT+ IT+ IN+ A+ BUTTER- DISH- AND+ KEPT+ FEEDING+ IT+ MOTHS AND- THINGS+ NASTY+ SO- I+ HATE+ LIKE- THE- WINTER+ THAT'S- WHEN+ THEY+ START+ TO+ COME-

"One", and other numbers, are particularly common in the above.

Rotation 9, cluster 6

84.896 WOO- WOO+ WOO+ WOO+ WOO+ WOO+ WOO+ WOO+ WOO+ WOO+
84.896 WOO- WOO+ WOO+ WOO+ WOO+ WOO+ WOO+ WOO+ WOO+ WOO+
67.129 HE'S- EX- FREEZE- JUMP+ AROUND+ JUMP- AROUND+ JUMP+ JUMP- JUMP- HI- HO+ HI+ HO+ IT'S+ TO- DO+ MY- HISTORY PROJECT- THAT'S- MY- FUCKING- HI- HO+ HI+ HO+ HI+
66.806 THAT+ LITTLE+ TINY- BLUE- FLOWER- AH- BUM+ BUM+ BUM+ BUM+ BUM+ BUM+ BUM+
55.950 I'M- HERE+ WOO+ WOO+ WOO+ WOO+ WOO+ WOO+ WOO+
36.504 TWINKLE- TWINKLE+ LITTLE+ STAR+ HOW+ I+
36.303 TWINKLE- TWINKLE+ LITTLE+ STAR+ HOW+ I+
33.019 OKAY+ LISTEN- LISTEN+ LISTEN+ A- BAR+ WITH- RAISINS+ MUESLI+ AND+ OATS+ CALLED- OAT+ BAR- YEAH+ THAT'S- MY- DESIGN A+ BAR+ WITH- NUTS+ AND+ RAISINS+ WHAT- CAN+ THAT- BE- CALLED+ MUM+ GOES+ NATURE-

What would normally be relatively unusual words appear in the above, often repeated.

Rotation 9, cluster 7

83.354 YOU- SCUM- OF+ THE+ EARTH+ DE+ DE+ DE+ DE+ DE+ DE+ DE+ DE+ YOU- SHOULD ONLY+ HAVE THAT- ATTITUDE+ IF- YOU- COME FROM A+ DEPRIVED+
64.638 I- KNOW- IT'S+ A+ STUPID+ PHRASE- BUT- YOU- LEARN BY+ YOUR- MISTAKES+ AND+ I'M+ TELLING- YOU TOO+ IT'S+ BLOODY- TRUE+ I+ KNOW- I'VE- LEARNT+ A+ FEW- THINGS BY+ HUGE+ MISTAKES+ YOU- DO- LEARN+ BY+ DAVE+ OBVIOUSLY DIDN'T+ KNOW- NO+ RIGHT- YEAH+ BUT- THERE'S+ SO- MUCH- DAMAGE+ TO+ YOU- YOU- SAW+ THAT+ FILM- YOU- SAW+ WHAT+ HAPPENED+ TO- THAT+ PERSON- IT'S+ AWFUL+ YEAH+ I'M- SAYING- THEY'RE+ BEING+ EDUCATED+ AREN'T- THEY+ THEY'RE+ BEING+ EDUCATED+ IF- THEY+ IF- THEY+ HAVEN'T+ LEARNT+ THEIR+ LESSONS- IF- THEY+ HAVEN'T+ IF- THEY+ HAVEN'T+ PAID+ ATTENTION+ IN- CLASS+ IT'S+ THEIR+ FAULT+ IT'S+ THEIR+ FAULT+ I KNOW- BUT YOU- SEE+ THEY'LL- REGRET+ IT+ LATER- BUT- THEY+ NEED- A- BIT+
46.182 YES+ THAT'S+ PARTICULARLY+ IMPORTANT+ IN- SYSTEMS+ LIKE- OURS+ WHERE- YOU- CAN DIVERT+ YOUR- TELEPHONE+ SO- EVEN+ THOUGH- YOU'RE+ CERTAIN- THAT+ YOU+ DIALLED+ THE+ RIGHT+ NUMBER+ YOU- COULD+ END+ UP+ ABSOLUTELY- ANYWHERE- BECAUSE- THE+ NUMBER+ YOU- DIALLED+ COULD- BE+ DIVERTED+ SOMEWHERE+ COMPLETELY- DIFFERENT+ SO- IT'S+ VERY- IMPORTANT+ WHEN+ YOU- ANSWER- THE TELEPHONE- TO+ SAY- WHO+ YOU- ARE+ THE+ OTHER- THING- THAT'S+ ANNOYING- ABOUT- THAT IS IT- THEN+ FORCES+ YOU- INTO A+ COMPLETELY+ USELESS- SMALL- CONVERSATION- SUCH+ AS+ IS- THAT- SO+ AND+ SO- AND+ THEY+ SAY- YES+ AND+ YOU- THEN+ FEEL- LIKE THEY+ SAY+ YES+ AS+ MUCH- AS TO+ SAY+

```
WELL- WHY- DIDN'T+ YOU- KNOW- THAT ANYWAY+ AND+ THEN-
YOU- FEEL- LIKE SAYING+ WELL+ WHY+ DIDN'T+ YOU- SAY- SO+
AND YOU- START OFF+ ON+ THE+ WRONG+ FOOT- ANYTHING-
  46.095 OH+ WELL- I+ KNOW- JUST- THOUGHT- I'D+ CALL+
YOU- ALRIGHT+ SO+ YOU- COULD+ YEAH+ YOU'VE+ GOT+ HIS-
TELEPHONE+ NUMBER+ ANYWAY- SO YOU- CAN- ALSO+ PHONE- HIM+
AT- THE SAME- TIME+ I+ JUST SAID LOOK+ YOU+ KNOW+ SILLY-
REALLY+ COS+ I- MEAN+ HE+ KNEW+ I- HAD+ A COUPLE+ OF+
PEOPLE+ YOU KNOW+ MONDAY+ AND- TUESDAY+ BEFORE+ MONDAY+
AND- TUESDAY+ AND+ YOU- KNOW YOU- GOT YOU NEED A- COUPLE+
OF+ PEOPLE+ AS- WELL+ SO- IF- YOU- DON'T- MIND+ COMING+
OVER+ FROM+ IT'S- ENTIRELY- UP- TO
```

"You", especially in combinations such as "you know" is especially common in these.

Rotation 9, cluster 8 :

```
 36.817 MM- WHAT- TO+ THINK+ OH+ I+ CAN'T+ THINK+ OF+
ANY+ WORDS+ OH- WE'RE+ GOING+ TO+ GO- OFF+ FOR- AN+
INTERVIEW+ NOW- ANYWAY+ MM- LOOKS+ VERY+
  30.830 MM- HAVE+ TO+ BE CAREFUL+ WHEN+ YOU'RE+ TALKING+
ABOUT+ THAT-
  30.054 EGG+ AND+ SAUSAGE+ WHAT+ KIND+ OF+ EGG+
  27.325 SEVEN+ FOUR+ SEVEN+ HUNDRED+ SEVEN+ FOUR+ SEVEN+
  26.878 MM- LEAVE- HIM+ A+ LITTLE- NOTE+ I+ THINK- WE'D+
BETTER+ GO+
  24.481 WHOSE+ WHOSE+ MOTHER+ DID+ YOU- SEE+ THIS+
MORNING+ MM-
  23.952 MM- PERHAPS- I+ OUGHT+ TO+ HAVE+ A+ GO+ AT+
  23.926 EXCUSE+ ME+ WHO'S- THAT+ MAKING- THAT+ NOISE+
  22.442 THEY+ DO+ THEY+ GO+ THEY GO+ AND- THEY+ BREAK+
INTO+ SHOPS+
  20.738 YOU+ FEEL+ BUT- WHEN+ YOU- GOT- IT+ ON+ RIGHT-
YOU+ FEEL+ SORT- OF LIKE+ MM- MM- MM- MM- MM- MM- MM-
AND+ YOU FEEL+ LIKE+ LOOK- AT- ME+ EXPENSIVE- EQUIPMENT+
IT'S+ PROBABLY+ ONLY+ WORTH+ ABOUT+ A+
  20.566 MM- YESTERDAY+ WHEN+ I+ REMEMBERED+ IT+ I+ JUST+
CRACKED+ UP+
```

"Mm" is particularly common in these turns.

So, in many cases, the production of a cluster using the lexically-based metric appears to be very heavily influenced by the presence or absence of a single specific word.

It was normally less easy to spot such trends in the clusters produced using the entropy-based metric, although some such patterns were observed.

e.g. "O" type experiments, using entropy-based metric with 10 clusters

Rotation 9, cluster 0

46.274 AND- HE WAS- PLAYING COMPUTER- WITH ME+ AND- I-
MEANT- TO- COME- UP TO YOU+ COS+ WHEN- THEY+ GO+ AWAY+
COS- THEY'RE- NOT- GOING+ AWAY+ THIS+ WEEKEND+ BUT+ WHEN
THEY GO+ AWAY+ YOU- SHOULD+ COME- DOWN+ COS+ THE- FLAT'S-
QUIET+ AND- I AIN'T GOT NOTHING+ TO DO SIT+ AND- PLAY-
MY+ COMPUTER- AND- IT WAS JUST SO- BRILLIANT+ I+ COULD
LEAVE+ EVERYTHING- LYING- ABOUT+ ON- THE- FLOOR+ AND-
NOTHING+ WOULD- GET BROKEN- I COULD LEAVE+ A+ MADONNA-
TAPE+ IN- THE- MIDDLE OF- THE ROOM- AND- NO ONE+ WOULD+
RIP- IT- I'D+ SIT+ AND+ WATCH+ IT+ YOU- KNOW+ JUST PUT+
IT+ THERE+ SIT+ AND- WATCH+ IT+ IT WAS BRILLIANT+ I+
COULD WATCH+ TELLY AND- ACTUALLY- HEAR+ WHAT JIM+ WAS+
SAYING+ IN- NEIGHBOURS+ IT WAS AMAZING+ I'VE NEVER I'VE+
NEVER+ EVER HEARD- JIM'S- VOICE

43.759 NO+ I WAS- THINKING+ OF HISTORY WHAT+ WHERE
THEY- ARE+ IN- HISTORY- SO+ YOU'VE NOW+ GOT+ TWO- FOUR+
SIX+ EIGHT TEN+ TWELVE+ FOURTEEN+ SIXTEEN+ EIGHTEEN+
TWENTY+ TWENTY+ ONE+ SO- FAR+ FOUR+ MORE+ THAT'S+
PUSHING- IT DAD'LL- LET+ TWENTY- FIVE I+ THINK+ AND-
EVERYBODY'S- GOT- TA+ BRING+ A PRESENT+ SO- THAT'S
TWENTY+ FIVE- PRESENTS- I- GET- TWENTY+ FOUR+ PRESENTS+
THAT'S- NOT+

42.346 YOU+ BITCH THE- MIKE AIN'T THE- MIKE'S- GONE
AGAIN+ TESTING TESTING+ ONE+ TWO+ THREE+ IT'S- GONE+
YEAH+ I'M+ GON NA OH+ TESTING- TESTING+ ONE+ TWO+ THREE+
YOU+ CAN HEAR+ YEAH+ IT'S+ ON+ THERE+ WHAT+ THE- FUCK+
ARE+ YOU+ DOING

40.492 RIGHT+ FOR- HOW+ MANY+ YEARS+ HAVE+ WE+ BEEN-
TOLD IT'S+ TAX- PAYERS'+ MONEY+ DO+ YOU+ REMEMBER-
MAGGIE- AND- THE- TAX- PAYERS'+ MONEY+ IT'S+ LIKE- THIS+
ANIMAL- SOMEWHERE- CALLED- THE TAX- PAYER- BUT IT CAME-
OUT OF+ THE WALL+ AS- IF+ WE+ WEREN'T+ ONE- OF- THEM+
AND- THAT+ WE- HAD+ TO LOOK+ AFTER+ THE TAX- PAYERS'+
MONEY+ WHAT ARE+ THEY+ DOING+ WITH+ MY+ MONEY+ NOW+
THEY'RE+ BRIBING- PEOPLE- LEFT- RIGHT+ AND+ CENTRE+ WITH+
IT+ I OBJECT- TO+ THAT MIND- YOU I SUPPOSE+ IF- THE-
OTHERS- WERE- IN THEY'D+ DO EXACTLY THE- SAME THING
WOULDN'T+

39.611 DO+ DO+ DO+ DO+ CHOO- DO+ DO+ DO+ DO+ CHOO- I
REALLY+ LIKE+ THAT- I PLAY+ IT+ ON+ MY+ BROTHER'S-
COMPUTER I'M+ ALWAYS BLOWING- THE- LITTLE+ DUCKS- UP-
THEY- GO+ THEY'RE GOING- LIKE+ THAT AND- YOU+ SHOOT-
THEM- AND- THEY- TURN- INTO- A- IT'S REALLY GOOD+ I-
LOVE+ THE- MUSIC+ IT GOES- DO+ DO DO+ DO+ CHOO- DO+ DO+
DO+ DO+ CHOO- DO+ DO+ DO+ DO+

The words "Do" and "Cos" and numbers are quite common in this cluster.

Rotation 9, cluster 3 :

60.327 WHAT- I'M- SAYING TO+ YOU+ NOW+ IS+ I+ WAS+ SAYING+ TO+ ARTHUR- THE- THE+ PETROL+ MONEY+ THAT- WE+ USED+ TO+ PUT- IN FOR- PETROL+ IT- COST+ US+ FIVE+ POUND+ A+ WEEK+ IN+ THE- WINTER+ BUT- IN+ THE SUMMER+ IT- COST+ US+ EIGHT+ POUND+ A+ WEEK+ SOMETIMES+

56.930 COS+ I'M+ IN BED- RIGHT- COS+ SOMETIMES+ I+ GO+ IN+ THE WEEK+ IT+ DEPENDS- HOW+ I- FEEL+ LIKE+ TONIGHT- I'LL- PROBABLY+ GO+ TO+ BED- ABOUT+ ABOUT+ HALF+ TEN+ OR+ ELEVEN+ COS- I'M+ GON- NA WATCH+ RUBY- WAX+ AND-

56.028 HE'S DOING+ IT AS+ A WEDDING+ PRESENT+ HE'S+ NOT+ CHARGING+ THEM+ ANYTHING+ LIKE+ THAT'S- A- I- SAYS+ WHAT+ CAN WE- GIVE+ THEM+ AS+ A+ WEDDING+ PRESENT+ HE+ SAYS- I'M+ GIVING+ ME+ DAUGHTER+ THAT'S- A- WEDDING+

51.858 HE+ JUST- WENT+ HELLO+ KATH- I- WENT+ RIGHT- YEAH+ YEAH+ COS- I- DO- ACTUALLY+ KNOW+ WHO+ YOU+ ARE+ I+ JUST- REMEMBER+ THE+ TIME+ HE+ CAME+ INTO- AND- HE+ HAD+ A- FAG+ AND+ HE+ SET- THE- ALARM+

44.540 ABOUT+ TEN+ PAST+ TWELVE+ STARTS- TO WORK+ OI- THINK OF A+ PROJECT+ FOR+ ME THAT- I+ CAN+ DO FOR+ THE NEXT SEVEN+ WEEKS+ WHAT+ CAN+ HE+ DO+ FOR RIGHT- DRAW+ LOADS- OF+ PEOPLE+ WITH- THEIR+ HEADS+ BEING- BLOWN- OFF- YEAH- THAT'S- A-

42.639 ME- ME+ MUM+ AND+ DAD SPLIT- UP+ JUST+ BEFORE+ DIDN'T+ THEY+ AND+ COULDN'T AFFORD+ IT+ SHE'S+ DOWN- TO+ HER+ TARGET+ WEIGHT- NOW+ SHE'S+ GOT+ TA- DO ANOTHER+ TWO- WEEKS+ AT+ AT+ AT+ SAME+

37.615 NO+ YOU- CAN'T+ YOU+ CAN'T+ RIGHT WE'RE+ GON+ NA+ START+ WITH- B+ THAT+ I'M+ GON- NA DO- IN+ AND+ THE+ NEXT- PERSON+ WHO'S+ STILL+ TALKING+ WHEN+ NO+ YOU'RE+ NOT IF+ GETTING+ OUT+ OF+ MY CLASS+ NOW+ I'M GOING- TO- AND+ THEN- GET INDIGESTION- AND+ THEN- AND+ AND+ ALL+ THAT+ I THINK+ YOU- CAN+ START OFF BY+ HAVING+

33.831 AND IN+ FACT+ IT+ STARTS- OFF+ WITH+ HIM+ IN- THE- GYM+ DOING+ HIS+ WORK+ OUT- AND- HE+ HAS+ THIS+ SONY- WALKMAN+

32.942 AND+ WELL+ OF+ COURSE+ THEN+ SOMEBODY+ SUGGESTED+ OH- WHAT- ABOUT+ SCOTTISH- COUNTRY+ DANCING+ COS+ THE SO+ I SAID+ OH- I'D+ LOVE+ SCOTTISH- COUNTRY+

32.043 WHERE+ ARE THOSE+ TAPES+ FOR THAT+ YOU- SAID+ I- COULD+ HAVE- ONLY+ ONE- THAT'S ALL+ RIGHT- I MEAN+ LIKE+ I+ ACTUALLY+ SORT- OF+ WORKED+ OUT- ABOUT+ ANOTHER+ TWO- OR+ THREE+ TAPES- YESTERDAY+ THE+ ONLY+ THING+ SEE- IS+ THAT- IT'S+ ALL+ MY- OWN+ TAPES- THEY'RE+ SUCH+ CRAP- ONES- AND+ LIKE THEY+ JUST- SOUND- REALLY+ BAD+ WHEN+ THEY'RE+ RECORDED- ON- IT'S LIKE+ THE- SAME- THING+ WITH

VIDEO- TAPES+ THEY+ ALL+ GET+ BUGGERED+ UP- SO+ I'VE-
GOT- TO SORT+ OF+ I'VE- GOT- TA- RECORD-
31.621 WELL- OURS+ IS+ SORT+ OF+ WELL DEREK+ I+ TOLD+
YOU NEW- NEW+ YEAR+ ON- THE+ SUNDAY+ HE+ WENT+ UP+ AND+
JUST+ THREW+ SOME- OF+ OURS+ OUT- AND+
31.511 BUT+ IT- ALWAYS+ SEEMS+ TO BE+ SOMETHING+ WE+
WANT+ SPECIAL+ EITHER+ A- SILLY- LITTLE+ COMEDY- AND+
THINGS+ LIKE+ THAT+ THEY+ WERE+ PERFECT- BUT+ ANYTHING-
WE+ REALLY+ WANT- IT+ ALWAYS+ SEEMS+ TO
30.498 WICKED+ YOU- KNOW+ WHAT+ I'VE- GOT- TA- BLOODY-
DO- YEAH+ TO- BRING THIS+ BACK- YEAH- THEY'VE- ASKED+ ME-
YEAH- TO- GET+ UP+ AT+ TEN+ THIRTY+ IN+ THE+ MORNING+
AND+ GO- TO- THE- SCHOOL+ ON- MONDAY+ WHEN+ I+ COULD- BE
LYING+ IN+ BED+ OH-
29.462 I DU+ N NO BUT+ SHE+ SAYS THAT+ SHE'S- YEARS+
AGO+ SHE+ USED+ TO+ PUT- AN+ ONION+ IN+ HER+ EAR- OR-
SOMETHING+ WHEN+

The above seem to include a lot of words related to time and dates, including numbers.

Rotation 9, cluster 4 :

52.878 YES+ WELL- YOU+ KNOW+ WENDY+ YOU+ COULD- COME+
AND+ AND- HAVE+ A+ SPOT+ OF+ LUNCH- IF+ YOU'D+ LIKE+ I+
MEAN+ YOU'RE+ ALWAYS+ SO- WELCOME+ TO+ COME+ AND+ HAVE+
ANYTHING+ YOU+ ONLY+ JUST- HAVE+ TO+ SAY+ WELL+ YOU+
KNOW+ WOULD- IT- SUIT- YOU+ FOR- ME- TO+ COME+
41.406 I+ DON'T+ KNOW THAT- ONE- OH+ YOU+ DO+ YOU'RE+
SO- BULLSHIT- JUST- TO+ THE- REAL- ONES+ WHATEVER+ THEY+
ARE+ OH- THEY+ GET- ALL+ THE+ I'M- TALKING+ I'M+ TALKING+
TALKING+ SEE+ I+ DID+ IT- IT'S- ME- JUST+ DO+ IT+ I+
JUST+ DID+ IT+ DO+ THE LONG- ONE+ AS-
40.998 NO+ THEY'RE+ NOT+ IGNORANT+ NOT- IGNORANT+ WHEN
THEY+ SAY- WELL+ THEY+ ARE+ IGNORANT+ COS- I+ WOULD-
NEVER+ SAY+ WHICH+ PART+ OF- FELLOW- SAID+ TO+ ME+ THE-
OTHER+ WEEK+ YOU+ DON'T LIVE+ HERE+ I+ SAID OH YES+ I+
37.791 I+ GOT- HER+ THERE- SHE- SAID+ SHE+ WAS IN+ A-
HURRY+ BECAUSE- SHE+ WAS- AFRAID+ TO+ LEAVE+ KEITH- TOO-
LONG+ I+ SAID+ WELL+ I'M+ IN+ A+ HURRY+ AS- WELL+ SO+ I+
GOT+ HER+ THERE- AS+ QUICK+ AS+ I+ COULD+
36.702 THEY+ SAY- THEY+ ARE+ GOING+ TO+ SHIP+ THEM+
WITH+ EVERY+ SYSTEM+ SEVEN+ YOU- GET+ THIS+ SO+ I+ WENT+
IN+ THERE+ AND+ I+ THOUGHT OH+ GOD+ I+ MEAN- I+ COULD+
IF+ YOU+ GAVE- ME+ IF+ I+ DID+ A- NICE- SKETCH- AND+
WORKED- ON+ IT+ IF+ I+ SPENT- A+ DAY- A-
33.491 ALRIGHT+ I+ MEAN+ NORMALLY+ I+ JUST KEEP+ GOING+
YOU+ KNOW+ AND+ SORT+ OF- DON'T+ BUT- I'VE+ BEEN- ACHING+
SO+ MUCH+ THAT+ IT'S+

28.954 I+ MEAN+ I'M+ SKINT+ AT- THE MOMENT I'VE+ I'VE+ I'VE+ HARDLY+ GOT+ ANY+ MONEY+ MONEY+ AT+ THE-

28.066 AND+ I+ OFFERED+ THEM+ A+ FIVER- TO+ PAY+ FOR+ THE+ VAN+ BUT- THEY+ DIDN'T+ TAKE+ IT IT- WAS- VERY- NICE+ VERY+ BUT+ YOU+ HAVE+ TO+ TAKE-

28.023 OH+ YEAH+ WELL- YES- WE'VE+ HAD- WE'VE+ PUT+ BEDDING+ PLANTS+ BEDDING+ PLANTS+

26.365 WHAT'S+ WRONG+ WITH+ THAT+ I+ CAN'T+ BARN+ DANCE+ BUT- I+ CAN+ HAVE- A+ JOLLY GOOD+ BASH+ AT+

26.103 STOP- A- MINUTE+ BEND- OVER+ I'VE- GOT TA+ WRITE+ DOWN+ WHO'S- TALKING+ TO- ME+ I+ JUST- GOT- TA+ WRITE+ DOWN+ WHO'S- TALKING+ TO- ME+

25.997 YEAH+ I+ CAN+ HEAR+ YOU+ LOUD- AND+ CLEAR+ COMING- THROUGH+

25.263 I+ HEAR+ YOU+ DIDN'T+ HAVE+ YOU- DIDN'T- ENJOY+ YOUR+ LIVER+ THE- OTHER DAY+ OX- LIVER+ OR- SOMETHING+ IS- THAT+

24.310 RIGHT- I'M+ JUST+ GOING+ TO+ DESTROY+ THIS- NOW+ ANDY- WHAT- DO+ WE+ NEED+ TO+

24.023 COULD+ YOU+ GIVE+ ME+ SOME+ IDEA+ OF+ HOW+ IT- HAS-

These are mostly quite short turns containing a high proportion of pronouns.

Rotation 9, cluster 6 :

48.420 RIGHT+ NOW+ YOU'RE- DOING- WHAT I- USED+ TO+ DO- WHAT+ I USED+ TO+ DO- WHEN+ I STARTED+ OKAY- YOU'RE- RUSHING+ I- STILL DO+ OKAY+ YOU'RE- YOU'RE+ RUSHING+ THROUGH+ YOU- ARE+ YOU'RE+ STARTING+ OFF+ SO+ YOU- GO+ THEN+ YOU- SUDDENLY+ REALIZE- I- KNOW- THIS+ YOU'RE-

44.081 J+ J+ WHAT'S- GOING+ ON WITH+ J+ J+ WHAT- DO- YOU- THINK+ DO I- LOOK+

43.846 YOU'RE- FIVE+ EIGHT+ YOU'RE- FIVE+ EIGHT+ YOU- STAND+ YOU- I- DON'T- THINK+ SAME+ SIZE- AS+ MY+ MUM+ I- DON'T- THINK+ SO SOMEHOW+ COS- MY+ MUMS+ FIVE+ EIGHT+ AND+ MY+ MUMS+ QUITE- TALL+ SHE'S NOT+ FIVE+

41.740 WELL- I- LIKE+ THE+ BLUE- AND+ YELLOW+ ONES- WELL- I'VE+ HAD+ ONE+ TWO+ THREE+ FOUR+ FIVE+ SIX+ CONVERSATIONS- OOH

40.981 NO- YEAH+ OH MY+ GOD+ OH+ MY+ GOD+ HE+ REALLY+ DOES+ OH- MY+ GOD+ THAT'S REALLY- WEIRD+ I THINK RICHARD'S+ HERE- OH- I- KNOW+ RICHARD'S- MUMS- CAR+ IT'S-

40.420 WELL- IT'S- IT'S- GON+ NA+ BE VERY- LATE+ THERE+ YOU'LL+ BE+ GETTING+ HER+ OUT+ OF- BED+ IF+ YOU'RE+ NOT- CAREFUL+ I- KNOW- SHE+ SAYS+ IT- DOESN'T+ MIND+ SHE+ SHE+ DOESN'T- MIND+ IT- DOESN'T- MATTER- BUT- OH+ DEAR- THERE'S- A+ PIECE+ HERE+ IN+ THE+ NEWSPAPER+ DID- YOU- SEE+

32.345 OH- YEAH+ THAT'LL- SUIT+ YOU RIGHT+ DOWN+ TO+ THE+ GROUND+ YOU'RE- A+ WEE+ THAT+ WAS+ OBVIOUSLY+ WHAT- I- WAS+ GON+ NA+ SAY+ AHA+ JUST- BECAUSE+ IT'S- YOUR- BIRTHDAY+ I'LL- LET+ YOU+ GET+ AWAY- WITH+

29.307 IF+ HE+ WERE+ TAKING+ SEVERAL- OF+ YOU- OUT+ WHY- DIDN'T+ YOU+ SAY+ TO+ ONE OF+ THE+ OTHER- GIRLS+ WHY- DON'T- YOU COME+ DID HE+ HAVE+ A+ PAIR- OF+

28.950 I- SUDDENLY+ WONDERED- IF+ I- WAS+ ALLOWED+ YEAH HE'S+ SUCH+ AN+ ARSEHOLE- BUT+ I- CAN'T- BELIEVE- WHEN+ HE+ CALLED+ YOU- A+ SLUT+

28.390 BUT IF+ I EVER+ HAVE+ IT- DONE+ AGAIN+ I'M- GOING+ TO BUY+ THOSE+ BRASS- THINGS+ AND+ IT- ONLY+ COSTS+ ABOUT+ FIVE+ POUNDS- MORE+ WHEREABOUTS+ DID+ YOU- GET+

25.275 OH- MY+ GOD+ WE+ DIDN'T+ KNOW- WHAT- HE+ WAS+ DOING- I+ MEAN+ IT- WOULD+ HAVE+ TO+ HAPPEN- TO- ME+ YEAH- THAT WAS+ SICK+ HONESTLY- IT- REALLY- WAS+ I- WAS+ LIKE- BIG+ FAT+

25.100 WHAT+ AH- I+ CAN'T FIND- MY+ YES I- AM+ SORRY+ I+ CAN'T- SMELL- THE+ GOOD- ONES+ I- DIDN'T+ KNOW- LIKE- THIS- M+ S+ FROM- M+

24.883 I- DON'T- KNOW+ SHE+ JUST+ SAID+ THAT+ HE+ WAS+ TWO+ FACED+ THEN+ AGAIN- I- THINK+ MOST+ OF+ THE+ PEOPLE+ HERE+ ARE+ I- DON'T- KNOW+ WHAT- YOU- CAN+ GET AWAY+

24.497 THEY'RE+ UP- TO THEIR+ EYES+ IN+ IT YOU CAN'T+ ASK+ THEM+ WHILE+ THEY'RE- THEY'RE IT'S- NOT-

These are again mostly rather short turns, where "do" & "don't", "did" & "didn't", "can" & "can't", etc. are quite common.

Rotation 9, cluster 7 :

87.721 HE+ WENT- UP+ AND- IT'S- FIRST- TIME+ HE'D+ SEEN+ HER+ FOR+ A+ WHILE+ AND+ SHE+ SAID+ SOMETHING+ ABOUT+ OH+ HE+ WAS+ SUPPOSED+ TO HAVE+ SOMETHING+ BUT HE+ GOT- TO- HEAR- THIS+ WELL- IT'S- GOT- NOTHING+ TO+ DO- WITH- ALL+ THE+ OTHERS+ COS ALEC- SAID+ SOMETHING+ TO+ ME+ ABOUT- MAGGIE+ I+ SAYS+ WELL+ WHAT AH+ I+ DON'T+ UNDERSTAND+ IT+ BECAUSE+ MARGARET+ ALEC- GOT- ON+ WELL+ WITH+ MARGARET+ MAYBE+ THEY+ BOTH KNEW+ THAT IS- I+ DON'T+

78.798 LIKE WHEN+ SHE- GETS+ THERE- SHE'S+ GON+ NA+ BE- PANICKING- ABOUT+ ONE- THING+ AND- ANOTHER+ SHE+ I- THINK+ WELL+ THEY+ SORT+ OF SAID+ THAT WHAT- HOPEFULLY+ SHE'LL+ LIKE+ BE- THERE FOR- A+ COUPLE+ OF DAYS+ AND+ SHE'LL+ NEVER+ THINK SHE'S+ BEEN+ ANYWHERE+ ELSE+ SHE'LL+ THINK+ SHE'S+ BEEN+ THERE+ ALL+ THE+ TIME+ BUT+ I+ DON'T THINK- SHE+ WILL+ I'LL+ TELL+

62.801 YEAH- SO- I+ SAID+ SO- IS- YOUR+ MUM+ AND+ DAD+ COMING+ DOWN- FROM+ BRISTOL+ THIS- WEEKEND+ SHE+ SAID+

```
NO+ MY+ FRIENDS+ COMING- DOWN- AND+ I+ SAID WELL+ DID+
YOU+ MANAGE+ TO- DO+ ANYTHING+ LAST- SATURDAY+ SHE+ SAID+
WE- WENT+ TO+ THE+ ISLE+ OF+
```

This cluster contains a lot of pronouns, notably a high proportion of third person feminine ones : "She", "her", "she's", "she'll", etc.

7.8 Summary

In this chapter, experiments have been described in which various language models for the dialogue material in the BNC were constructed in a manner based on clustering the available data.

The first approach, in which complete dialogues were clustered, constructed a trigram model for each of 10 clusters. These were then combined by interpolation, together with an ordinary trigram model, to give a "mixture model". The weightings of the components for the mixture model were found both for a "static" model (where the optimal weights were computed based on the whole set of data reserved for this purpose) which was then applied to the full test dataset, and for an adaptive model (where the weights were adapted on the basis of the most-recently seen data from the test set), with the model being adapted after every 10% of the available test data. In both cases, the mixture models showed very little improvement in perplexity over the ordinary trigram model. This suggests that clustering of whole dialogues for use in a mixture model of the kind described here is of little value in modelling the BNC dialogue data, or at least that any updating of the weights of such a model should be done much more frequently – each 10% chunk of the test data will contain many dialogues from several different sources and there is little (if any) reason to suppose that dialogues from distinct sources should be connected in any way.

The second approach involved clustering pairs of successive, relatively short, dialogue turns using a lexically-motivated metric or an entropy-based metric. It was expected that knowledge of the nature of the first turn of a pair – i.e. to which cluster that first turn would be allocated – would be useful in predicting the content of the second turn of the same pair. Three different "valid" approaches to cluster choice, plus one "oracle" (cheating) approach were used. In any one experiment, M clusters

were found for the training set and a language model produced for each. Individually, each cluster model was interpolated with a simple trigram model. In testing, these interpolated models gave a modest improvement in perplexity over the ordinary trigram model. Not surprisingly, the "oracle" method out-performed the others, giving a bound on the best improvement which could possibly be obtained by this type of approach. It was found that the perplexity obtained using a given clustering strategy improved as the number of clusters used was increased (at least up as far as 40 clusters). However, for a fixed number of clusters, there was no significant change in perplexity if the size of the lexicon used in the language modelling was varied.

These results suggest that clustering the data can be of values in the statistical modelling of dialogue, particularly if the dependencies and similarities we are hoping to model are of the appropriate type.

Chapter 8 Discussion, Conclusions and Suggestions for Further Work

8.1 Discussion and Conclusions

As noted in chapter 3, the majority of language models used in modern automatic speech recognition systems have been trained on text material or transcriptions of news broadcasts (or similar) and the acoustic models used have typically be trained on read speech. However, it was argued that read speech is somewhat different to spontaneous speech, and (even once transcribed), spontaneous speech – and dialogue in particular – is quite different from written text material. It would seem much more appropriate to train both language and acoustic models on material as close as possible to the type of speech and language to which they are to be applied.

The work presented in this thesis, based on the large body of dialogue material within the BNC, has provided evidence in support of the hypothesis put forward in chapter 3 that dialogue material is rather different from text in several ways.

At the most basic level, the variety of vocabulary is considerably smaller in the dialogue material than in the text, and (with the exception of restricted vocabularies of up to 9 words), a lexicon consisting of only the most common N words (chosen appropriate to the material currently being studied) accounts for a higher proportion of the BNC dialogue material then the corresponding N word text lexicon does for the BNC text data. The smaller lexicon size for dialogue data leads to simplification of the problem of constructing language models. The dialogue vocabulary can be regarded as a closed system – if we can be sure that our vocabulary includes all the distinct words which might be encountered within both the training and test data, then there is no concern about how to deal with out-of-vocabulary words.

A wide variety of lengths of sentences, dialogue turns , dialogues (defined here to allow only two speakers per dialogue) and "conversation files" (a record of a set of one or more dialogues occurring in sequence at one location and event) were observed within the BNC dialogue data. In some cases, the so-called dialogue files contained

several speakers and/or lengthy sections where the speaker did not change. This prompted the sub-division of the data files into dialogues (with only two speakers each) and the creation of the "Dialogue, Reduced Turns" (DRT) dataset, comprising pairs of relatively short dialogue turns which were considered to be more typical of spontaneous highly interactive dialogue and to emphasise short-range, structural features rather than aspects heavily dependent on a longer scale, such as the topic of the conversation.

Separate simple trigram language models were constructed for the "ordinary" dialogue data (taken directly from the appropriate BNC files), for the DRT dataset and for samples of BNC text data of various sizes. These models were then applied to "reserved data" of the appropriate type and their perplexities with respect to that data calculated. The perplexities of the models for "ordinary dialogue" and DRT data were quite similar, but those for the text models – even those trained on significantly more data – were much higher. Although the perplexities of the text models decreased as the size of the training set was increased, extrapolation of the trend suggested that approximately 500 million words of text training material would be required to achieve results comparable with those trained on 7 million words of dialogue. These results ("improvement, but with diminishing returns" as the size of the training set is increased) are broadly in line with those of Lamel et al (2002) and Moore (2003) on the amount of data required to achieve various targets of Word Error Rates (WERs) for an automatic speech recognition system performing a specified task. The sensitivity of the model to the type of data on which it is trained and tested is consistent with the observations of Rosenfeld (2000b) and Young (2000).

As noted below, one consequence of the trigram language models for dialogue having much lower perplexities than those for text is that it is more difficult to gain large improvements in perplexity, by refining the language models through the use of more advanced techniques, for dialogue data than it is for text material.

Attempts to allow more "adaptability" of the language model to the nature of the conversation of current interest were made using cache and trigger pair models. The cache models, when interpolated with the basic trigram model, gave a noticeable improvement for both "ordinary dialogue" and DRT material. This appears to be

largely due to repetitions of words and phrases, both by the same speaker and by the other speaker – features which are particularly common in dialogue, used for speech repairs, for emphasis and for confirmation or clarification purposes - in addition to the kind of repetitions which tend to occur in text material. Even use of a very small cache (just a few words) proved beneficial. Thus, the cache model, despite being quite simple conceptually and easy to implement, is really quite effective in the modelling of dialogue. This has analogies with psycholinguistic approaches to dialogue, where small "caches" within short term memory are believed to play a role in the processing of conversation by humans (e.g. Walker 1996, Purver, Ginzburg & Healey 2002).

In contrast to this, the benefits obtained using the trigger pair based models on dialogue were less than those obtained using a cache. Previous work on text data (e.g. Rosenfeld 1996) had suggested that a model based on word trigger pairs could be quite powerful and of major benefit in improving the adaptability of language models. However, it was noted in this study that very few triggers tended to be active at any point. This was partly due to the short nature of typical sentences and turns in dialogue and the relatively small window sizes which, consequentially, could be used. When no triggers are currently active, the exponential probability model using trigger pairs effectively uses a "default" probability - namely (1/W), where W is the number of possible words (possibly of a "restricted" vocabulary, where the most common and/or least common words are excluded). This may give a specific word a very unrealistic probability – either too high or two low – both with respect to its normal frequency of occurrence and the current context. These factors are believed to be a major difficulty with regard to a trigger-based model being of major benefit in the modelling of highly interactive dialogues with relatively short turns. Furthermore, training these trigger models proved very expensive in terms of the computational time and memory required. Thus, such trigger models – at least of the "word trigger pair" type investigated here – would appear to be of limited value when modelling short dialogue turns.

As recommended by Rosenfeld (1996), potential word trigger pairs were selected on the basis of their mutual information with respect to training data. It was hoped that some trigger pairs found for dialogue data in this way would be linguistically notable in some way. Unless restrictions were imposed on either the words which could act

as triggers, or as target words, or both, for dialogue material the majority of such trigger pairs yielding a relatively high mutual information were extremely common words. A model based on such trigger pairs would not be expected to yield much benefit over a simple trigram model, where the probabilities of very common words are generally well-estimated. However, some interesting trigger pairs, showing reasonably high mutual information, were found in the dialogue material. Some of these were trigger pairs which did not appear in the corresponding list for text material – indeed, some included rather "colloquial" words which would not be common in ordinary written text material. On the other hand, some trigger pairs highly-ranked in the both the lists for dialogue and for text data, although interesting, did appear to be quirks of some of the training material rather than widely useful trigger pairs for the general case.

It was hoped that training a trigger-based model on the DRT data might yield some trigger pairs more related to structural features of dialogue rather then just pairs of lexically-related words relevant to the current theme and context of the conversation. To some extent, some such trigger pairs were found which had high mutual information. For example, certain related pronouns which might share their "resolution" (i.e. person/object to which they refer) were found to act as triggers for each other : e.g. "she", "she's", "she'd" and "her" tended to trigger each other. However, there was no strong evidence for certain other expected features of dialogue, such as a turn featuring "why" being followed by one featuring "because" (indicative of a question being followed by an answer), within the lists of most promising trigger pairs.

Another approach to allowing the language model to adapt to the material of current interest was to cluster the turn pairs according to a similarity criterion – "similar" turns being put in the same cluster. Two different similarity (or difference) metrics were used – one "lexically" based : on how distinctive any given word was to a particular group of documents compared with any other group, the other based on the perplexity (or entropy) of the resulting cluster-based models – and four different rules for dealing with pairs of turns were employed. Some promising results were obtained. One approach – an "oracle" method, where information within the second turn was used to choose the cluster language model appropriate for that turn – showed a

significant reduction in perplexity of up to about 13% when comparing an interpolated trigram-cluster model with the trigram model alone. However, it should be noted that this "oracle" method is something of a "cheat" as far as a realistic automatic dialogue system would be concerned, since it relies on information within a given turn to "predict" things about the same turn. Nevertheless, similar "oracle" approaches, e.g. to the modelling of sequences of dialogue acts, have been used by previous authors (e.g. Stolke et al, 2000). The results from the use of such an "oracle" approach can be considered as an upper bound to the improvements in perplexity which could be aspired to through the use of turn clustering. The other three approaches to clustering the data – which could be genuinely incorporated into a real automatic dialogue system - gave more modest improvements of up to 3% with respect to the baseline trigram model when interpolated with it.

This case of the "oracle" experiments also produced an interesting property – a "dustbin" effect. The clusters produced varied greatly in size – often, one cluster contained as much of the data as all the other clusters put together. The perplexity of the language model corresponding to this largest cluster was usually much higher than that of any other. It would appear that the effect of applying this method is to sweep all turn pairs which are highly unpredictable into this single large "dustbin" cluster – hence its very high perplexity, whilst the other, more predictable (and hence easier to model) turns are divided into relatively homogeneous clusters in a useful manner. The two different metrics gave similar results – the lexically-based metric gave slightly lower perplexities when just a few clusters were allowed, whereas the entropy-based metric gave better perplexity values when larger numbers of clusters were used. In both cases, the perplexities of interpolated trigram-cluster models decreased as the number of clusters used was increased, indicating that finer discrimination between types of turn were possible when more clusters were employed. However, in terms of computational speed, the entropy-based metric gave significantly better performance.

Some interesting linguistic features of the resulting clusters were observed. The content of the clusters of turns produced using the lexically-based metric did show certainly similarities between the individual turns, but in most cases these were at a very simple level – particular individual words, or words taken from a small group (e.g. numbers or "wh- question" words) might occur in most or all or the turns within

a given cluster. A similar trend – although rather less pronounced – was found amongst some of the clusters produced by the entropy-based metric. However, there was little or no evidence of the clusters of turns produced by either of these methods showing any real consistency of topic.

This study has shown that improvements over simple trigram models can be made to statistical language models for dialogue using techniques which allow the language model being used to adapt to the material currently of interest. However, the relative benefits of these adaptive methods – cache, trigger and cluster-based techniques – are not necessarily the same as for modelling written text material where, for example, use of a model based on word trigger pairs would be expected to yield more benefit than the use of a relatively small cache. This study has shown that – at least when comparing the text and dialogue data within the BNC – there are features of dialogue material which are significantly different from ordinary text. As outlined above, the differences in relative utility of cache, trigger and cluster based models for text and dialogue data are believed to be due to such features.

Some of the results of this study – notably some of the trigger pairs showing highest mutual information with respect to the training corpus, both in the cases of dialogue and text material – illustrate that, in its present form, the BNC is far from ideal as a training corpus for language models intended for applications in automatic speech recognition or speech understanding systems. On the one hand, the range of material – in terms of the sources and topics covered – in the BNC is too broad to be appropriate for training a model aimed at a specific type of application such as news transcription, ticket booking or travel enquiries. On the other hand, the material contained in the BNC – even when, say, we only consider the dialogue data – is not sufficiently representative of the full range of modern British spoken English in terms of topic of conversation to be the ideal source of training material for a "general purpose" speech recognition or dialogue system. For either of these aims, the construction of new, large corpora of training material – either focusing on a specific domain, or attempting to be as general and representative as possible – would be desirable.

8.2 Suggestions for Further Work

This project has yielded several interesting questions as yet unanswered. It has provided further evidence for the hypothesis that dialogue material is significantly different in nature from written text, and that this should be taken account of in statistical language modelling. This is of particular importance when constructing and training language models for automatic spoken dialogue systems, for applications such as automated enquiry services.

Most automatic speech recognition systems have acoustic models and language models which were both trained on read text material. This study has illustrated that, at the language modelling level, dialogue differs significantly from text. Indeed, it was observed (see section 4.6) that even simple language models trained and tested on dialogue data from the BNC had much lower perplexities than equivalent models trained and tested on BNC text material. Therefore, it would seem appropriate that the language model for a dialogue system should be trained on dialogue material. It would also appear likely that other, more acoustic, features of speech – such as intonation and co-articulation – may differ between read speech and spontaneous dialogue. Some studies of this nature have already been carried out (e.g. Taylor et al 1998, King 1998) .

As noted in section 8.1 above, the "quality" of a language or acoustic model (measured in terms of having a low perplexity or word error rate) tends to improve as the quantity of material used to train it is increased. Bearing in mind the above observation – that models for dialogue applications should be trained on dialogue material, it would be desirable that much larger corpora of dialogue material were available. The "dialogue" portion of the BNC contains approximately 7.7 million words, but a substantial part of that consists of long "pseudo-monologues" where the speaker changes only occasionally or, in some cases, one speaker accounts for almost all the words spoken. Furthermore, quite a large number of the files were recorded in controlled situations such as school lessons where the interaction between the speakers is not typical of more general dialogue. The vocabulary used in such

situations (e.g. in a chemistry or mathematics lesson) may also be highly atypical of everyday conversation.

During the study of the word trigger pairs produced from the BNC dialogue material, it was noted that some pairs showing relatively high mutual information were very unusual – for example, the name "Collymore" occurred both as a self-trigger and in association with the word "football" (probably due to frequent occurrence in BNC file HMN – a football commentary from "The Central Match"). Similarly, descriptions of mathematical notation and relatively obscure chemicals occurred in the list of promising trigger pairs. Likewise, when no restrictions were placed on the trigger or target words, some of the highest-ranked trigger pairs for text data were rather surprising (with "award" and "zero" as extremely common target words). None of these examples would be expected to be particularly beneficial to the modelling of "typical" English dialogue (or text). This illustrates how sensitive the language model – in this case a trigger-based one- can be to the nature of the material on which it is trained relative to the material to which it is to be applied, consistent with the observations of Rosenfeld (2000b) and Young (2000). To model "typical" British English dialogue successfully, our training data should be genuinely representative of British English dialogue ! However, production of such a "genuinely representative" training corpus is likely to be difficult at best. For example, how should we judge whether the spontaneous dialogue material which had been collected was "genuinely representative" of typical modern British English conversation, or whether the range of sources was "sufficiently broad" ? For more specific applications, training material should both be as extensive as practicable and appropriate to the application for which the model is to be used. It would be interesting to investigate how the results of this present study compared with those for a corresponding set of experiments on a corpus of dialogue material on a more restricted (but more uniform) range of topics.

Although other large corpora containing dialogue material do exist, most of these are of American English and some only contain dialogues on a specific topic or relating to a particular task. The recording, annotation and mark-up of a much larger corpus of spontaneous dialogue material – preferably with acoustic information such as intonation included - would be of benefit to both language and acoustic modelling of dialogue. However, the production of such a corpus is likely to be both highly labour-

intensive and expensive. Nevertheless, some of the more restricted corpora could be of value in studies like the one suggested in the preceding paragraph above.

The use of cache models in dialogue modelling proved to be a simple but relatively effective way of making the language model adapt to the material of current interest, with even a very small cache proving useful. Experiments using a decaying cache – where word probabilities are not just calculated according to their presence in (or absence from) the cache, or their frequency within the cache, but according to how far back in the history they occurred – as used by Clarkson & Robinson (1997), might prove of benefit, particularly since dialogue turns tend to be relatively short and word repetitions on a relatively short distance scale (in words) common. Some evidence from semantic studies on requests for clarifications in dialogue (Purver, Ginsburg & Healey 2002) suggests that repetitions (for purposes of clarification or acknowledgement) are particularly common at short distances within dialogue. This would support the use of such a decaying cache.

The application of clustering techniques to the dialogue material yielded interesting results, even if the most successful approach (the "oracle" method) was really something of a cheat. Nevertheless, this illustrated that some significant benefit was in principle obtainable from the use of language models based on clusters of dialogue turns or pairs of dialogue turns. This approach may well be worth further investigation, particularly from a point of view of a stochastic state model, or "cluster transition model", with the clusters as the states – i.e. if it is known that the current turn belong to cluster A, construct a model in order that the probabilities of the following turn belonging to each of clusters A, B, C, etc. can be estimated. This could possibly incorporated into a "mixture of clusters" model, so that words in the following turn can be predicted according to probabilities obtained according to each cluster model, weighted by the individual cluster probabilities given by the "cluster transition model". This idea shares some features with the dialogue move models used by some previous authors (e.g. Stolke et al 2000, Reithinger 1996, Wahlster 2000) and also with some aspects of trigger models. Perhaps trigger pairs based on clusters – where the presence of a word or phrase from one cluster triggers another "target cluster" (rather than a specific target word) – might show more success than a model based on word trigger pairs. Possibly even a single word or short phrase could be a

useful trigger to predict the cluster to which the next turn should belong. Alternative strategies for clustering turns should also be explored. For example, some approach to turn clustering as yet untried might lead to the automatic classification of turns into dialogue acts, or even suggest alternative categories to some of those currently in general use.

Finally, there has traditionally been a rather strict division between the "Statistical Language Modelling" and "Natural Language Processing" (knowledge-based methods using syntactic parsing and/or semantics) approaches to the computational modelling of English. It would be hoped that some benefit could be obtained by attempting to produce a hybrid approach, trying to capitalise of the strongest features of both methodologies. For example, a statistical approach to classification or clustering of dialogue turns as "dialogue acts" could be combined with a knowledge-based approach for analysing transitions between such acts.

References

Alexandersson, J., Maier, E. & Reithinger, N. (1995) "A Robust and Efficient Three-Layered Dialogue Component for a Speech-to-Speech Translation System", *Proceedings of the European Chapter of the Association for Computational Linguistics*, Dublin, pp 188-193. (Also Verbmobil Report 50, DFKI, Saarbrücken)

Alexandersson, J. & Reithinger, N. (1997) "Learning Dialogue Structures from a Corpus", *Proceedings of Eurospeech '97*, Rhodes, Greece, pp 2231-2235.

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N., Schmitz, B. & Siegel, M. (1998) "Dialogue Acts in VERBMOBIL-2" (Second edition), Report 226, DFKI Saarbrücken, July 1998.

Alexandersson, J., Engel, R., Kipp, M., Koch, S., Küssner, U., Reithinger, N., & Stede, M. (2000) "Modeling Negotiation Dialogs", in "Vermobil : Foundations of Speech-to-Speech Translation", Editor : W. Wahlster, Springer Verlag, Berlin, pp 441-451

Allen , J.F. & Perrault, C.R. (1980) "Analysing Intention in Dialogues", *Artificial Intelligence*, Vol. 15, Number 3, pp 143-178

Allen, J.F., Byron, D.K., Dzikovska, M., Ferguson, G., Galescu, L. & Stent, A. (2001) "Towards Conversational Human-Computer Interaction", *AI Magazine*, Vol. 22, No. 4, pp 27-37.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R. (1991). "The HCRC Map Task Corpus", *Language and Speech,* **34**, pp. 351-366.

Austin, J. (1962) "*How to Do Things with Words*", Harvard University Press, Massachusetts, USA

Baddeley, A. (1986) "*Working Memory*", Oxford University Press, Oxford, U.K.

Bahl, L., Jelinek, F. & Mercer, R. (1983) "A Maximum Likelihood Approach to Continuous Speech Recognition", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **5**, pp 179-190

Baker, J.K. (1975) "The Dragon System - an Overview", *IEEE Transactions on Acoustics, Speech & Signal Processing*, **23** (1), pp 24-29

Bard, E.G., Sotillo, C., Anderson, A.H., Taylor, M.M. (1995) "The DCIEM Map Task Corpus", *Proceedings of the ESCA-NATO Tutorial and Workshop on Speech Under Stress*, Lisbon, Portugal.

Bard, E.G., Sotillo, C., Anderson, A.H., Thompson, H.S. & Taylor, M.M. (1996) "The DCIEM Map Task Corpus : Spontaneous Dialogue Under Sleep Deprivation and Drug Treatment", Speech Communication, Vol. 20, pp 71-84.

Beeferman, D. , Berger, A. & Lafferty, J. (1997) "A Model of Lexical Attraction and Repulsion", *Proceedings of the ACL-EACL Joint Conference '97*, pp 373-380.

Bellegarda, J.R. (2000) "Exploiting Latent Semantic Information in Statistical Language Modelling", *Proceedings of the IEEE*, 88 (8), pp 1279-1296

Berger, A.L., Della Pietra, S.A. & Della Pietra, V.J. (1996) "A Maximum Entrpy Approach to Natural Language Processing", *Computational Linguistics*, Vol. 22, pp 1-36

Bod, R. (2000) "Combining Semantic and Syntactic Structure for Language Modelling", *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China.*

Bodner, G.E. & Masson, M.E.J. (2003) "Beyond Binary Judgements : Prime Validty Modulates Masked Repetition Priming in the Naming Task", Memory & Cognition (in press). http://web.uvic.ca/psyc/masson/Bodner4.pdf (downloaded 3 February 2004)

Brill, E., Florian, R., Henderson, J.C. & Mangu, L. (1998) "Beyond N-Grams : Can Linguistic Sophistication Improve Language Modelling ?", *Proceedings of COLING/ACL 1998 Conference, Montreal, Canada*, Vol. I, pp 186-190.

Burch, Gull, S.F. & Skilling, J. (1983) *Computer Vision, Graphics & Image Processing*, **23**, pp 111-124

Burnard, L. (1995) "Users' Reference Guide for the British National Corpus", Oxford University Computing Services, Oxford, UK.

Cahn, J.E. & Brennan, S.E. (1999) "A Psychological Model of Grounding and Repair in Dialog", Proceedings of the Fall AAAI Symposium on Psychological Models of Communication in Collaborative Systems, Sea Cliff, Massachusetts, USA (Nov. 5-7 1999), pp 25-33

Cameron, D. (2001) "*Working with Spoken Discourse*" (Sage Publications, London)

Canavan, A. & Zipperlen, G. (1997) "CallFriend American English-Non-Southern Dialect" by A. Canavan and G. Zipperlen, Linguistic Data Consortium, USA http://www.ldc.upenn.edu/Catalog/LDC96S46.html (downloaded 24 October 2001)

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. and Anderson, A. (1997). "The Reliability of a Dialogue Structure Coding Scheme", *Computational Linguistics*, Vol. 23, pp. 13-31

Carter, D. (1994a) "Improving Language Models by Clustering Training Sentences", Proceedings of the 4th Association for Computational Linguistics Conference on Applied Natural Language Processing, Stuttgart, October 1994.

Carter, D. (1994b) "Improving Language Models by Clustering Training Sentences", Technical Report, SRI International, Cambridge, U.K.

Chen, S.F. & Goodman, J. (1999) "An Empirical Study of Smoothing Techniques for Language Modelling", *Computer Speech & Language*, **13** , pp 359-393

Chen, S.F. & Rosenfeld, R. (2000) "A Survey of Smoothing Techniques for ME Models", *IEEE Transactions on Speech & Audio Processing*, **8** (1), pp 37- 50

Chomsky, N. (1957) "Syntactic Structures", p 15
(Reprinted 1976, Mouton, The Hague)

Chu-Carroll, J. (1999) "Form-Based Reasoning for Mixed-Initiative Dialogue Management in Information-Query Systems", Proceedings of Eurospeech '99, Budapest, Vol. 4, pp 1519-1522

Clark, H.H. (1994) "Managing Problems in Speaking", *Speech Communication*, Vol. 15, Numbers 3-4, pp 243-250

Clark, H.H. (1996) "Using Language", Cambridge University Press, Cambridge, U.K.

Clark & Haviland (1977) "Comprehension and the Given-New Contract", in *Discourse Production and Comprehension* (Editor : R.O. Freedle), Ablex Publishing Corporation, Norwood, New Jersey, USA.

Clark, H.H. & Marshall, C.R. (1981) "Definite Reference and Mutual Knowledge", in *Elements of Discourse Analysis* (Editors : A.K. Joshi, B.L. Webber & I.A. Sag), Cambridge University Press, Cambridge, U.K.

Clark H.H. & Schraefer, E.F. (1987) "Collaborating on Contributions to Conversations", *Language and Cognitive Processes*, Vol. 2, pp 1-23

Clark, H.H. & Schaefer, E.F. (1989) "Contributing to Discourse", *Cognitive Science*, Vol. 13, pp 259-294

Clark, H.H. & Brennan, S.E. (1991) "Grounding in Communication", in Perspectives on Socially Shard Cognition (Editors L.B. Resnick, J. Levine & S.D. Teasley), pp 127-149, APA, Washington DC, USA.

Clarkson, P.R. & Robinson, A.J. (1997) "Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache", *Proceedings of the International Conference on Acoustics, Speech & Signal Processing*, Vol.2, pp 799-802

Clarkson, P.R. & Robinson, A.J. (1998) "The Applicability of Adaptative Language Modelling for the Broadcast News Task", *Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP'98*, Sydney, Australia

Clarkson, P. & Rosenfeld, R. (1997) "Statistical Language Modeling using the CMU-Cambridge Toolkit", Proceedings of Eurospeech '97, Vol. 5, 2707-2710

Clarkson, P.R. (1999) "Adaptation of Statistical Language Models for Automatic Speech Recognition", PhD Thesis, University of Cambridge, U.K.

Cohen, P.R. & Perrault, C.R. (1979) "Elements of a Plan-Based Theory of Speech Acts", Cognitive Science, Vol. 3, Number 3, pp 179-212

Cohen, P.R., Levesque, H.J., Nunes, J.H.T & Oviatt, S.L. (1990)"Task-Oriented Dialogue as a Consequence of Joint Activity", *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, Nagoya, Japan, November 1990, pp 203-208

Dahlbäck, N. & Jönsson, A. (1999) "Knowledge Sources In Spoken Dialogue Systems", *Proceedings of Eurospeech '99, Budapest*, Vol. 4, pp 1523-1526.

Darroch, J.N. & Ratcliff, D. (1972) "Generalised Iterative Scaling for Log-Linear Models", *Annals of Mathematical Statistics*, **43**, pp 1470-1480

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm" , Journal of the Royal Statistical Society, **39** (1), pp1-38

Denes, P. (1959) "The Design and Operation of the Mechanical Speech Recogniser at University College London", *Journal of the British Institution of Radio Engineers*, **19**, pp 219-229

Ebert, C., Lappin, S, Gregory, H. & Nicolov, N. (2001) "Generating Full Paraphrases of Fragments in a Dialogue Interpretation System", Proceedings of the Second SIGDial Workshop of Discourse & Dialogue, Aarlborg, Denmark, pp 58-67.

Ehara, T., Ogura, K. & Morimoto, T. (1990) "ATR Dialogue Database", *Proceedings of ICSLP'90*, pp 1093-1096

Fernández, R. & Ginzburg, J. (2002) "Non-Sentential Utterances : A Corpus Study", Traitment Automatique des Languages : Dialogue, Vol. 43, No. 2, pp 13-42

Fischler, I. & Bloom, P.A. (1980) "Rapid Processing of the Meaning of Sentences", *Memory & Cognition*, Vol. 8, pp 216-225

Fletcher, C. (1994) "Levels of Representation in Memory for Discourse", in *Handbook of Psycholinguistics* (Editor : M. Gernsbacher), Academic Press, New York

Fry, D.B. (1959) "Theoretical Aspects of Mechanical Speech Recognition", *Journal of the British Institution of Radio Engineers*, **19**, pp 211-219

Fry, D.B. & Denes, P. (1955)"Experiments in Mechanical Speech Recognition", In *Information Theory* (Butterworth & Co., London), pp 206-212

Ginzburg, J. (1997) "On Some Semantic Consequences of Turn Taking", *Proceedings of the 11th Amsterdam Colloquium*, (Editors : P. Dekker , M. Stokhof & Y.Venema), ILLC, Amsterdam, pp 145-150

Ginzburg, J. (1998) "Clarifying Utterances", *Proceedings of the 2nd Twente Workshop on the Formal Semantics & Pragmatics of Dialogue and 13th Twente Workshop on Language Technology*, Editors : J. Hulstijn & A. Nijholt, Twente, the Netherlands

Ginzburg, J. (2001a) "The ROSSINI Project" http://www.dcs.kcl.ac.uk/research/groups/nlp/rossini.html (downloaded 25th May 2003)

Ginzburg, J. (2001b) "Semantics and Interaction in Dialogue", CSLI Publications, Stanford, USA and Cambridge University Press.

Ginzburg, J. (2003) "PROFILE : Processing and Resolution Of Fragments In dialoguE" http://www.dcs.kcl.ac.uk/staff/ginzburg/profile.html (downloaded 25th May 2003)

Ginzburg, J., Gregory, H. & Lappin, S. (2001a) "SHARDS : Fragment Resolution in Dialogue", Proceedings of the 4th International Conference on Computational Semantics, Tilburg, The Netherlands, pp 156-172.

Ginzburg, J. & Sag, I. (2001) "Interrogative Investigations : the form, meaning and use of English Interrogatives", CSLI Lecture Notes Series, University of Chicago Press, Chicago, USA

Ginzburg, J., Sag, I. & Purver, M. (2001b) "Integrating Conversational Move Types in the Grammar of Conversation", Proceedings of BI-DIALOG 2001: 5th Workshop on the Semantics & Pragmatics of Dialogue, Bielefeld, Germany (Editors : P. Kühnlein, H. Rieser & H. Zeevat)

Godfrey, J.J., Holliman, E.C. & McDaniel, J. (1992) "SWITCHBOARD : Telephone Speech Corpus for Research & Development", *Proceedings of the International Conference on Acoustics, Speech & Signal Processing*, Vol. 1, pp 517-520

Godfrey, J.J. & Holliman, E. (1997) "Switchboard-1 Release 2", Linguistic Data Consortium, USA http://www.morph.ldc.upenn.edu/Catalogue/LDC97S62.html (downloaded 27 September 2001)

Gorin, A.L., Riccardi, G. & Wright, J.H. (1997) "How May I Help You ?", *Speech Communication*, **23**, pp 113-127

Gorin,A.L., Wright, J.H, Riccardi, G, Abella, A. & Alonso, T. (2001) "Semantic Information Processing of Spoken Language", Workshop on Innovation in Speech Processing (WISP 2001), Stratford-upon-Avon, UK, April 2001, *Proceedings of the Institute of Acoustics*, **23** (3), pp 63-70

Gregory, H. (2001) "The HPSG Dialogue Project" http://semantics.phil.kcl.ac.uk/dialogue/about.html (downloaded 30th May 2003)

Greig, D.M. (1980) "Optimisation" (Longman Group, London) pp 6-21

Grosz, B.J. (1977) "The Representation and Use of Focus in Dialogue Understanding", PhD Thesis, University of California, Berkley, U.S.A.

Grosz, B., Joshi, A.K. & Weinstein, S. (1995) "Centering : A Framework for Modelling the Local Coherence of Discourse", *Computational Linguistics*, Vol 21, No. 2, pp 203-225

Grosz, B. & Sidner, C.L. (1986) "Attention, Intentions and the Structure of Discourse", *Computational Linguistics*, Vol. 12, No. 3, pp 175-204

Grosz, B. & Sidner, C.L. (1998) "Lost Intuitions and Forgotten Intentions", in *Centering in Discourse*, Editors : M.A. Walker, A. Joshi & E. Prince, Oxford University Press, Oxford, U.K., pp 39-51

Gull, S.F. (1988) "Bayesian Inductive Inference and Maximum Entropy", in *Maximum Entropy and Bayesian Methods in Science and Engineering, Vol. 1 : Foundations*, editors G.J. Erikson & C.R. Smith (Kluwer Academic Publishers, Dordrecht, Boston & London), pp 53-74

Gull, S.F. & Skilling, J. (1984) "Maximum Entropy Method in Image Processing", *IEE Proceedings*, **131F**, pp 646-659

Halliday, M.A.K. & Hasan, R. (1976) "*Cohesion in English*", Longman Publishing, London, U.K.

Harris, Z. (1952) "Discourse Analysis", *Language*, Vol. 28, pp 1-30

HCRC (2001) "HCRC Map Task Corpus Annotations 1.0", Human Communications Research Centre, University of Edinburgh & University of Glasgow, http://www.hcrc.ed.ac.ac/maptask (version downloaded 20 September 2001)

He, Y. & Young, S.J. (2003) "Hidden Vector State Model for Hierarchical Semantic Parsing", Submitted to *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, Hong Kong, Apr. 6-10, 2003.

Heeman, P.A. & Hirst, G. (1995) "Collaborating on Referring Expressions", Computational Linguistics, Vol. 21, Number 3.

Hochberg, M., Rennals, S. & Robinson, A. (1995) "*ABBOT* : The CUED Hybrid Connectionist-HMM Large Vocabulary Recognition System", *Proceedings of the Language Technology Workshop, Austin, Texas, January 1995* (Morgan Kaufman Publishing)

Holcomb, P.J. (1993) "Semantic Priming and Stimulus Degradation : Implications for the Role of N400 in Language Processing", *Psychophysiology*, Vol. 30, pp 47-61

Huckvale, M. (1998) "Opportunities for Re-Convergence of Engineering and Cognitive Science Accounts of Spoken Word Recognition", *Proceedings of the Institute of Acoustics Conference on Speech & Hearing, Windermere, November1998*

Huckvale, M.A. & Hunter, G.J.A. (2001) "Learning on the Job : The Application of Machine Learning within the Speech Recognition Decoder", Workshop on Innovation in Speech Processing (WISP 2001), Stratford-upon-Avon, UK, April 2001, *Proceedings of the Institute of Acoustics*, **23** (3), pp 71-79

Hymes, D. (1972a) "On Communicative Competence", in *Sociolingusitics* (Editors : J.B. Pride & J. Holmes) pp 269-293, Penguin Books, Harmondsworth, U.K.

Hymes, D. (1972b) "Models of the Interaction of Language and Social Life", in Directions in Sociolinguistics (Editors : J.J. Gumperz & D. Hymes), pp 35-71, Holt, Rinehart & Winston, New York.

Iyer, R.M. & Ostendorf, M. (1999) "Modeling Long Distance Dependence in Language : Topic Mixtures Versus Dynamic Cache Models", *IEEE Transactions on Speech & Audio Processing*, **7** (1), pp 30-39.

Jaynes, E.T. (1988) "How Does the Brain Do Plausible Reasoning ?", in *Maximum Entropy and Bayesian Methods in Science and Engineering, Vol. 1 : Foundations*, editors G.J. Erikson & C.R. Smith (Kluwer Academic Publishers, Dordrecht, Boston & London), pp 1-24

Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M. & Quantz, J.J. (1995) "Dialogue Acts in Verbmobil" (First edition), VM-Report 65, DFKI Saarbrücken, April 1995

Jelinek, F. (1976) "Continuous Speech Recognition by Statistical Methods", *Proceedings of the IEEE*, **64** (4) pp 532-556

Jelinek, F. (1990)"Self-Organized Language Modelling for Speech Recognition", In Waibel, A. & Lee, K-F. (ed.) *Readings in Speech Recognition*, (Morgan Kaufmann, San Matteo, California), pp 450-506

Jelinek (1991) "Up From Trigrams !", *Proceedings of Eurospeech'91, Genova, Italy*, September 1991, Vol. 3, pp 1037-1040.

Jelinek, F., Merialdo, B., Roukos, S. & Strauss, M. (1991) "A Dynamic Language Model for Speech Recognition", *Proceedings of the DARPA Workshop on Speech & Natural Language*, February 1991, pp 293-295.

Jokinen, K. Hurtig, T., Hynnä, K., Kanto, K., Kaipainen, M. & Kerminen, A. (2001) "Self-Organising Dialogue Management", Proceedings of the 2nd workshop on Natural Language Processing and Neural Networks, Tokyo, pp 77-84.

Jokinen, K., Kerminen, A., Kaipainen, M., Jauhiainen, T., Wilcock, G., Kuusisto, J. & Lagus, K. (2002) "Adaptive Dialogue Systems – Interaction with Interact", Proceedinggs of the 3rd SIGDial Workshop on Discourse & Dialogue (ACL-02), Philadelphia, USA, July 2002, pp 64-73

Jurafsky, D., Bates, R., Coccaro, N., Martin, R. Mateer, M., Ries, K., Shrivberg, E., Stolcke, A., Taylor, P. & Van Ess-Dykema, C. (1997) "Automatic Detection of Discourse Structure for Speech Recognition & Understanding", *Proceedings of IEEE Workshop on Speech Recognition & Understanding*, Santa Barbara, California, December 1997 pp 88-95

Jurafsky, D., Bates, R., Coccaro, N., Martin, R. Mateer, M., Ries, K., Shrivberg, E., Stolcke, A., Taylor, P. & Van Ess-Dykema, C. (1998) "Switchboard Discourse Language Modeling Project – Final Report", Research Note 30, Johns Hopkins University, Baltimore, USA, January 1998

Katz, S.M. (1987) "Estimation of Probabilities from Sparse Data for the Langiuage Model Component of a Speech Recogniser", *IEEE Transactions on Acoustics, Speech ans Signal Processing*, 35 (3) 400-401

King, S.A. (1998) "Using Information Above the Word Level for Automatic Speech Recognition", PhD Thesis, University of Edinburgh, U.K.

Kingsbury, P., Strassel, S., McLemore, C. & McIntyre, R. (1999) "CallHome American English Transcripts", Linguistic Data Consortium, USA http://www.ldc.upenn.edu/Catalog/LDC97T14.html (downloaded 23 October 2001)

Kita, K., Fukui, Y, Nagata, M. & Morimoto, T. (1996) "Automatic Acquisition of Probabilistic Dialogue Models", *Proceedings of the International Conference on Spoken Language Processing* (ICSLP '96), Philadelphia, Editors : H.T. Bunnell & W. Idsardi, Vol. 1, pp 196-199

Klatt, D.H. (1977) "Review of the ARPA Speech Understanding Project", *Journal of the Acoustical Society of America*, 62 (6), pp 1324-1366

Kohonen, T. (1982) "Analysis of a Simple Self-Organising Process", *Biological Cybernetics*, Vol. 44, No. 2, pp 135-140

Kohonen, T. (1988) "The 'Neural' Phonetic Typewriter", *IEEE Computer*, March 1988, pp 11-21

Kohonen, T. (2001) "*Self-Organising Maps*", 3rd Edition (1st Edition 1995), Springer Verlag, Berlin

Kuhn, R. & De Mori, R. (1990) "A Cache-Based Natural Language Model for Speech Recognition", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 12 (6), pp 570 - 583, Corrections in *IEEE Trans. PAMI*, Vol. 14 (6), pp 691-692

Kurimo, M. & Lagus, K. (2002) "An Efficiently Focusing Large Vocabulary Language Model", *Proceedings of the International Conference on Artificial Neural Networks* (ICANN '02), Spain, August 2002.

Labov, W. (1972a) "The Study of Language in its Social Context", in *Language and Social Context*, Editor : P.P. Giglioli, pp 283-307, Penguin Books, Harmondsworth, U.K.

Labov, W. (1972b) "The Transformation of Experience in Narrative Syntax", in *Language in the Inner City*, pp 354-396, University of Pennsylvania Press, Philadelphia, USA.

Labov, W. (1972c) "*Sociolinguistic Patterns*", University of Pennsylvania Press, Philadelphia, USA.

Labov, W. & Fanshel, D. (1977) "*Therapeutic Discourse: Therapy as Conversation*", Academic Press, New York.

Labov, W. & Waletzsky, J. (1967) "Narrative Analysis", in *Essays on the Verbal and Visual Arts*, Editor : J. Helm, University of Washington Press, Seattle, USA, pp 12-44

Lagus, K. & Kuusisto, J. (2002) "Topic Identification in Natural Lngusge Dialogues Using Neural Networks", Proceedings of the 3rd SIGDial Workshop on Discourse & Dialogue (ACL-02), Philadelphia, USA, July 2002, pp 95-102
Lamel, L, Gauvain, J-L, & Adda, G. (2002) "Unsupervised Acoustic Model Training", *Proceedings of the International Conference on Acoustics, Speech & Signal Processing* (ICASSP '02), Vol. 1, pp 877-880.

Laplace, P.S. (1843) "Exposition de la Theorie des Chances et des Probabilites" (Paris). English Translation by Truscott-Emory "A Philosophical Essay on Probabilities" (Dover Publications, New York, 1951), p 196

Lappin, S. & Leass, H.J. (1994) "An Algorithm for Pronominal Anaphora Resolution", *Computational Linguistics*, Vol. 20, Number 4, pp 535-561

Lappin, S. & Gregory, H. (1997) "A Computational Model of Ellipsis", *Proceedings of the Conference on Formal Grammar*, ESSLLI, Aix-en-Provence, 1997 (Editors : G-J Kruiff, G. Morrill, and R. Oehrle)

Larsson, S. & Traum, D.(2000) "Information State and Dialogue Management in the TRINDI Dialogue Move Engine Toolkit", Natural Language Engineering, Vol. 6, Numbers 3-4, pp 323-340. See also the TRINDI website : http://www.ling.gu.se/projekt/trindi

Lau, R., Rosenfeld, R. & Roukos, S. (1993) "Trigger-Based Language Models : A Maximum Entropy Approach", *Proceedings of the International Conference on Acoustics, Speech & Signal Processing (ICASSP'93)*, Vol. 2, pp 45-48.

LDC (2001) "Bramshill Corpus", Linguistic Data Consortium, USA http://www.ldc.upenn.edu/Catalog/LDC94S20.html (downloaded 12 October 2001)

Lea, W.A. (1980) "The Value of Speech Recognition Systems"
In *Trends in Speech Recognition*, Chapter 1, pp 3-18
(Prentice Hall, New York)

Lippmann, R.P. (1997) "Speech Recognition by Machines and Humans", *Speech Communication*, **22**, 1-15.

Litman, D.J. & Allen, J.F. (1990) "Discourse Processing and Commonsense Plans", in *Intentions in Communication* (Editors : P.R. Cohen, J. Morgan & M.E. Pollack), pp 365-388, M.I. T. Press, Cambridge, Massachusetts, USA.

Lowerre, B. & Reddy, R. (1980) "The *Harpy* Speech Understanding System", in *Trends in Speech Recogntion*, Ed. W. Lea (Prentice Hall)

Mast, M., Kompe, R., Harbeck, S., Kiessling, A., Niemann, H., Nöth, E., Schukat-Talmazzini, E.G. & Warnke, V. (1996) "Dialog Act Classification with the Help of Prosody", *Proceedings of the International Conference on Spoken Language Processing* (ICSLP '96), Philadelphia, Editors : H.T. Bunnell & W. Idsardi, Vol. 3, pp1732-1735

Mateer, M. & Iyer, R. (1996) "Modeling Conversational Speech for Speech Recognition", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, USA, May 1996.

Meyer, D.E. & Schvaneveldt, R.W. (1971) "Facilitation in Recognizing Pairs of Words : Evidence of a Dependence Between Retrieval Operations", Journal of Experimental Psychology, Vol. 90, pp 227-234.

Miller, G.A. (1956) "The Magical Number Seven, Plus or Minus Two : Some Limits on our Capacity for Processing Information", *Psychological Review*, Vol. 99, No. 3, pp 440-466

Mitkov, R., Lappin, S. & Boguraev, B. (2001) "Introduction to the Special Issue on Computational Anaphora Resolution", *Computational Linguistics*, Vol. 27, Number 4, pp 473-478

Moore, R.K. (2003) "A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners", *Proceedings of Eurospeech 2003*, Geneva. September 2003 (to appear).

Nagata, M. (1992) "Using Pragmatics to Rule-Out Recognition Errors in Cooperative Task-Oriented Dialogues", *Proceedings of ICSLP '92*, Banff, Canada, Vol. 1, pp 647-650.

Nagata, M. & Morimoto, T. (1993) "An Experimental Statistical Dialogue Model to Predict the Speech Act Type of the Next Utterance", *Proceedings of the International Symposium on Spoken Dialogue* (ISSD-93), Tokyo, Japan, pp 83-86.

Nagata, M. & Morimoto, T. (1994) "First Steps Towards Statistical Modeling of Dialogue to Predict the Speech Act Type of the Next Utterance", Speech Communication, Vol. 15, pp 193-203.

Nelder, J.A. & Meade, R. (1965) "A Simplex Method for Function Minimisation", *The Computer Journal*, Vol. 7, pp 308-313

Ney, H., Essen, U. & Kneser (1994) "On Structuring Probabilistic Dependencies in Stochastic Language Modelling", *Computer Speech and Language*, Vol. 8, pp 1-38

Niesler, T.R. & Woodland, P.C. (1996) "A Variable Length Category-Based N-gram Language Model", *Proceedings of the International Conference on Acoustics, Speech & Signal Processing (ICASSP'96),* Vol. 1, pp 164-167

Nofsinger, R.E. (1991) "Everyday Conversation", Sage Publications, London, UK & Newbury Park, California, USA

Oviatt, S.L. & Cohen, P.R (1991) "Discourse Structure and Performance Efficiency in Interactive and Noninteractive Spoken Modalities", *Computer Speech & Language*, Vol. 5, pp 297

Paul, D.B. & Baker, J.M. (1992) "The Design for the Wall Street Journal-based CSR Corpus", *Proceedings of the DARPA Workshop on Speech & Natural Language*, pp 357-362

Pereira, F. (2000) "Formal Grammar and Information Theory : Together Again ?", *Philosophical Transactions of the Royal Society of London : Mathematical, Physical & Engineering Sciences*, **358** (1769), pp 1239-1253.

Perrault, C.R. & Allen, J.F. (1980) "A Plan-Based Analysis of Indirect Speech Acts", *American Journal of Computational Linguistics*, Vol. 6, Number 3, pp 167-182

Pierce, J.R. (1969) "Whither Speech Recognition ?", *Journal of the Acoustical Society of America*, **46**, pp 1049-1051.

Potamianos, A, Riccardi, G. & Narayanan, S. (1999) "Categorical Understanding Using Statistical N-gram Models", *Proceedings of Eurospeech '99, Budapest*, Vol. 5, pp 2027-2030

Power, R. (1979) "The Organisation of Purposeful Dialogues", *Linguistics*, Vol. 17, pp 107-152

Press,W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P (1992) "Numerical Recipes in C : The Art of Scientific Computing" (Second Edition), Cambridge University Press, pp 408-412

Purver, M., Ginzburg, J. & Healey, P. (2001) "On the Means of Clarification in Dialogue", Proceedings of the Second Association for Computational Linguistics SIGDial Workshop on Discourse & Dialogue, September 2001, pp 116-125

Purver, M., Ginzburg, J. & Healy, P. (2002) "On the Means for Clarification in Dialogue", in Current & New Directions in Discourse & Dialogue (Editors : R. Smiths & J. van Kuppevelt), Chapter 1, Kluwer Academic Publishers.

Rabiner, L.R. & Juang, B.H. (1986) "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine*, Vol. 3 (1), pp 4-16, January 1986

Ratnaparkhi, A. (1997) "A Simple Introduction to Maximum Entropy Models for Natural Language Processing", *University of Pennsylvania Institute for Research in Cognitive Science, Report 97-08.*

Reddy, D.R. (1976) "Speech Recognition by Machine : A Review", *Proceedings of the IEEE*, 64 (4), pp 502-531

Reithinger, N. (1995) "Some Experiments in Speech Act Prediction", *Proceedings of the AAAI'95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.* (Also Vermobil Report 49, DFKI, Saarbrücken, December 1994)

Reithinger, N., Engel, R., Kipp, M. & Klesen, M. (1996) "Predicting Dialogue Acts for a Speech-to-Speech Translation System", *Proceedings of the International Conference on Spoken Language Processing* (ICSLP '96), Philadelphia, Editors : H.T. Bunnell & W. Idsardi, Vol. 2, pp 654-657

Reithinger, N. & Klesen, M. (1997) "Dialogue Act Classification Using Language Models", *Proceedings of the 5th European Conference on Speech Communication & Technology* (Eurospeech '97), Rhodes, Greece, Vol. 4, pp 2235-2238.

Robertson, S.E. & Spärck Jones, K. (1997) "Simple Proven Approaches to Text Retreival", Technical Report TR 356, University of Cambridge Computer Laboratory, U.K. (http://www.cl.cam.ac.uk/TechReports/TRIndex.html)

Rosenfeld, R. (1994) "Adaptive Statistical Language Modelling : A Maximum Entropy Approach", PhD Thesis, Carnegie Mellon University, Pittsburg, P.A., U.S.A. (also available as Carnegie Mellon University School of Computer Science Technical Report CMU-CS-94-138)

Rosenfeld, R. (1996) "A Maximum Entropy Approach to Adaptive Statistical Language Modelling", *Computer Speech & Language*, Vol.10, 187-228

Rosenfeld, R. (2000a) "Incorporating Linguistic Structure into Statistical Language Modelling", *Philosophical Transactions of the Royal Society of London A*, 358, 1311-1324

Rosenfeld, R. (2000b) "Two Decades of Statistical Language Modelling : Where do we go from here ?" *Proceedings of the IEEE*, Vol. 88 (8) pp 1270-1278

Rosenfeld, R. & Huang, X (1992) "Improvements in Stochastic Language Modeling", in *Proceedings of the DARPA Workshop on Speech and Natural Language, February 1992* (Morgan-Kaufmann), pp 107-111

Sacks, H. (1972) "On the Analyzability of Stories by Children" in *Directions in Sociolinguistics* (Editors J.J. Gumperz & D. Hymes) Holt, Rinehart Winston, New York

Samuel, K., Carberry, S. & Vijay-Shanker, K. (1998) "Dialogue Act Tagging with Transformation-Based Learning", *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th Annual Conference on Computational Linguistics*, Montreal, Canada, Vol. 2, pp. 1150-1156
(See also Sandra Carberry, Ken Samuel, K. Vijay-Shanker, and Andrew Wilson. "Randomized Rule Selection in Transformation-Based Learning: A Comparative Study". *Natural Language Engineering,* 7(2), pp. 99-116, 2001.)

Scheffler, K. & Young, S. (2000) "Probabilistic Simulation of Human-Machine Dialogues", Proceedings of ICASSP 2000, Istanbul, Turkey, Vol. 2, pp 1217-1220

Schiffrin, D. (1994) "*Approaches to Discourse*", Blackwell Press, Oxford, U.K.

Searle, J. (1969) "*Speech Acts*", Cambridge University Press, Cambridge, U.K.

Sekine, S. (1994) "Automatic Sublanguage Identification for a New Text", *Second Annual Workshop on Very Large Corpora, Kyoto, Japan*, pp 109-120

Shannon, C.F. (1951) "Prediction and Entropy of Printed English Text", *Bell System Technical Journal*, **30**, pp 50-64

Siu, M. & Ostendorf, M. (2000) "Variable N-grams and Extensions for Conversational Speech Language Modeling", *IEEE Transactions on Speech and Audio Processing*, **8** (1), pp 63-75

Stanovich, K.E. & West, R.F. (1979) "Mechanisms of Sentence Context Effects in Reading : Automatic Activation and Conscious Attention", *Memory and Cognition*, Vol. 7, pp 77-85

Stolke, A, Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., Meteer, M. (2000) "Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech", *Computational Linguistics*, Vol. 26, No. 3., pp 339-374

STP (1996) (Switchboard Transcription Project) "Switchboard Transcription System" http://www.icsi.berkeley.edu/real/stp/description.html , downloaded 8 October 2001

Taylor, P., Shimodaira, H., Isard, S., King, S. & Kowto, J. (1996) "Using Prosodic Information to Constrain Language Models for Spoken Dialogue", *Proceedings of the International Conference on Spoken Language Processing* (ICSLP '96), Philadelphia, USA.

Taylor, P., King, S., Isard, S. & Wright, H. (1998), "Intonation and Dialogue Context as Constraints for Speech Recognition", Language & Speech, Vol. 41, nos. 3-4, pp 493-512. (489-508 ?)

Tillmann, B. & Bigand, E. (2003) "A Comparative Review of Priming Effects in Language and Music" http://olfac.univ-lyon1.fr/unite/equipe-02/tillmann/download/Tillmann_CNSPL.pdf , downloaded 9 December 2003

Traum, D.R. (1994) "A Computational Theory of Grounding in Natural Language Conversation", PhD Thesis, University of Rochester, New York, USA.

Traum, D.R. & Allen, J.F. (1992) "A 'Speech Acts' Approach to Grounding in Conversation", *Proceedings of ICSLP '92*, pp 137-140, October 1992

Traum, D.R. & Hinkleman, E.A. (1992) "Conversation Acts in Task-Oriented Spoken Dialogue", Computational Intelligence, Vol. 8, Number 3, pp 575-599

Traum, D.R. & Dillenbourg, P. (1996) "Miscommunication in Multi-Modal Collaboration", AAAI Workshop on Detecting, Repairing and Preventing Human-Machine Miscommunication (August 1996), pp 37-46.

Traum, D.R. & Heeman, P.A. (1996) "Utterance Units and Grounding in Spoken Dialogue", *Proceedings of ICSLP '96*, October 1996

Van de Waijer, J. (2001) "The importance of Single-Word Utterances for Early Word Recognition", *Proceedings of ELA 2001*, Lyon, France

Wahlster, W. (1993) "Verbmobil – Translation of Face-to-Face Dialogs", Technical Report, German Research Centre for Artificial Intelligence (DFKI) and *Proceedings of ATR Workshop on Speech Translation* (IWST'93), Kobe, Japan, November 1993.

Wahlster, W. (2000) (Editor) "*Verbmobil : Foundations of Speech-to-Speech Translation*", Springer Verlag, Berlin.

Waibel, A. & Lee, K-F. (1990a) "Why Study Speech Recognition ?"
In Waibel, A. & Lee, K-F. (ed.) *Readings in Speech Recognition*, (Morgan Kaufmann, San Matteo, California), Chapter 1, pp 1-5.

Waibel, A. & Lee, K-F. (1990b "Language Processing for Speech Recognition"
In Waibel, A. & Lee, K-F. (ed.) *Readings in Speech Recognition*, (Morgan Kaufmann, San Matteo, California), Chapter 8, pp 447-449.

Walker, M.A. (1992) "Redundancy in Collaborative Discourse", *Proceedings of COLING-92*, Nantes, France, August 1992.

Walker, M.A. (1993) "Informational Redundancy and Resource Bounds in Dialogue" Ph.D. Thesis, University of Pennsylvania, Philadelphia, USA (IRCS Report 93-45)

Walker, M.A. (1996) "Limited Attention and Discourse Structure", *Computational Linguistics*, Vol. 22, No.2, pp 255-264.

Walker, M.A. (1998) "Centering, Anaphora Resolution and Discourse Structure", in *Centering in Discourse*, Editors : M.A. Walker, A.K. Joshi & E.F. Prince, Oxford University Press, Oxford, U.K.

Walker, M.A.& Whittaker, S. (1990) "Mixed Initiative in Dialogue : An Investigation into Discourse Segmentation", Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL'90), Pittsburgh, Pennsylvania, USA, pp 70-76.

Williams, J. D. & Young, S. (2003) "Using Wizard-of-Oz Simulations to Bootstrap Reinforcement-Learning-Based Dialogue Management Systems", to appear in Proceedings of the SIGDial-ACL Annual Meeting, Sapporo, Japan, July 2003

Witten, I.H. & Bell, T.C. (1991) "The Zero-Frequency Problem : Estimating the Probabilities of Novel Events in Adaptive Text Compression", *IEEE Transactions on Information Theory*, 37 (4), pp 1085- 1094

Wright, H. (1998) "Automatic Utterance Type Detection Using Supersegmental Features", *Proceedings of the International Conference on Spoken Language Processing* (ICSLP '98), Sydney, Editors : R.H. Mannell & J. Robert-Ribes, Vol. 4, pp 1403-1406

Wright, H, Poesio, M. & Isard, S. (1999) "Using High Level Dialogue Information for Dialogue Act Recognition Using Prosodic Features", *Proceedings of the ESCA Tutorial & Research Workshop on Dialogue and Prosody*, pp 139-143.

Wright, H. (2000) "Modelling Prosodic and Dialogue Information for Automatic Speech Recognition", PhD Thesis, University of Edinburgh, U.K.

Young, S (1996) "Large Vocabulary Continuous Speech Recognition : A Review", *IEEE Signal Processing Magazine*, Vol. 13 (5), pp 45-57.

Young, S. (2000) "Probabilistic Methods in Spoken Dialogue Systems", *Philosphical Transactions of the Royal Society (London)*, Series A, Vol. 358, No. 1769, pp 1389-1402

Young, S. (2002) "Talking to Machines (Statistically Speaking)", *Proceedings of ICSLP 2002*, Denver, Colorado, USA.

Zhang, R., Black, E. & Finch, A. (1999) "Using Detailed Linguistic Structure in Language Modelling", *Proceedings of Eurospeech '99, Budapest*, pp 1815-1817

Appendix A

Some Further Examples of Turn Pairs Showing Large Increases in Probability Through Use of a Cache Model

From Test Set 0

54.366 [OKAY ALLOTROPES OF CARBON A CARBONATE PLUS AN ACID GIVES] AN+ ACID+ CARBONATE+ PLUS+ AN+ ACID+
36.489 [ONE FROM ELIZABETH AND ONE FROM RON] ONE+ FROM+ RON+ AND+ ONE+ FROM+ ELIZABETH+
36.433 [HARD RETURN AND SOFT RETURNS] HARD+ RETURN+ SOFT+ RETURN+
34.401 [WHERE'S TAB OH] TAB+ TAB+ TAB+
32.922 [BUT THERE'S ONE SCORE WHICH UNAMBIGUOUSLY CALLS IT IN C MINOR ON THE AND IT STARTS IN C MINOR SO] IT+ IT+ STARTS+ IN+ C+ MINOR+ IT+ STARTS+
31.786 [SHIFT F SEVEN] SHIFT+ SHIFT+ F+ SEVEN+
31.678 [EIGHT AND WELL FOUR TWOS TWO FOURS AND FOUR TWOS] TWO+ FOURS+ FOUR+ TWOS+ EIGHT+
31.101 [RIGHT AND THAT'S A GENERAL REACTION THAT HAPPENS WITH VIRTUALLY ANY ACID AND ANY ALKALI] ACID+ AND+ ANY+ ALKALI+
31.085 [ONE ONE THIRD JUST WRITE DOWN ONE THIRD ADD ONE SIXTH] ONE+ THIRD+ ADD+ ONE+ SIXTH+
28.648 [TAKE AWAY ONE TWELFTH OKAY JUST WRITE THAT DOWN THAT YOU'VE GOT TO TAKE AWAY THE ONE TWELFTH] ONE+ TWELFTH+ ONE+ TWELFTH+
27.603 [AND DAVID'S GON NA CONVALESCE] DAVID'S+ GON+ NA- CONVALESCE+
27.447 [EAST HERTS YEAH] YEAH+ EAST+ HERTS+
27.426 [CLUB BAR LICENCE ALAS] CLUB+ BAR+ LICENCE+
27.235 [SO IT'S TWO TIME WHAT ABOUT NINE EIGHT OR NINE FOUR NINE FOUR WE'D BETTER DO HADN'T WE] NINE+ FOUR+ IT'S+ NINE+ IT'S+ NINE+
26.937 [THE SONG SOUNDED BEAUTIFUL AND LONELY] BEAUTIFUL+ AND+ LONELY+ THE+ SONG+
26.935 [IT'S A BOGGIN CRASH] IT'S+ A+ BOGGIN+ CRASH+
26.836 [VERY WELL ON THAT NITRATES SULPHATES AND WHAT ELSE ANY OTHER HATES THAT YOU'VE HEARD OF] NITRATES+ SULPHATES+
26.327 [THE CURRENT IN THAT RESISTOR] THE+ CURRENT+ IN+ THAT+ RESISTOR+
25.943 [OH YES OOPS SORRY YOU YOU WRITE IT P U N C] P+ U+ N+ C+ P+ U+ N+ C+
25.943 [IT WAS CREEPING SUBURBIA THEN IT WASN'T] CREEPING+ SUBURBIA+
25.942 [THIS IS PURELY STAFF COSTS] PURELY+ STAFF+ COSTS+
25.831 [RIGHT SO YOU LOAD YOUR FILE F ONE F ONE] F+ ONE+ F+ ONE+
25.655 [ESSO'S THE TIGER] ESSO'S+ THE+ TIGER+
25.121 [SHE'S NO MORE MEDICINE LEFT] SHE'S+ NO+ MEDICINE+ LEFT+
24.662 [I WANT TO KNOW WHAT WORK WITH UNEMPLOYE ED IS] UNEMPLOYE+ ED+

24.559 [YEAH THERE IS THORLEY SAINSBURY'S] THORLEY+ SAINSBURY'S+
24.553 [EXACTLY IT'LL BE COS WE LOOKED AT THIS LAST WEEK AS YOU SAID AT THE END WHAT HAPPENS DRIPPING ACID ONTO ONTO CHIPS NOW THE THINGS TO KNOW ABOUT ACIDS BASES AND SALTS A METAL PLUS AN ACID WHAT HAPPENS] A+ METAL+ AND- PLUS+ AN+ ACID+
24.479 [WE GOT THEIRS WERE HIGHER THAN OURS] THEIRS+ WERE+ HIGHER+ THAN- OURS+
24.128 [AND SO SULPHUR TRIOXIDE ADD WATER MAKES SULPHURIC ACID] SULPHURIC+ ACID+
23.826 [IT MUST HAVE BEEN GOOD IT MUST HAVE BEEN SOME SORT OF CARBONATE AND THE SALT THAT WAS FORMED WAS FROM THE HYDROCHLORIC ACID WAS CALCIUM CHLORIDE SO IT MUST HAVE BEEN] CALCIUM+ CARBONATE+
23.811 [SO YOU'D NAME THIS AS BUTANE IN OTHER WORDS YOU'RE SAYING IT'S A BUTANE CHAIN YOU TAKE OFF THE E YOU WILL ADD O L AND IF THERE ARE POSITIONAL ISOMERS POSSIBLE YOU HAVE TO INDICATE THE POSITION ONE O L BUTANE ONE L ONE O L BUTANE ONE O L] BUTANE+ ONE+ O+ L+
23.740 [AND THEN A THREE DASH DIE] THREE+ DASH+ DIE+
22.792 [IT GETS BIGGER RIGHT SO YOU'VE TAKEN THAT DOWN A SEMITONE TO TO MAKE IT MINOR TO MAKE IT DIMINISHED YOU JUST TAKE IT DOWN ANOTHER SEMITONE] TAKE+ IT- DOWN+ ANOTHER+ SEMITONE+
22.656 [KILOWATT HOURS] KILOWATT+ HOURS+
22.608 [AND IT LIBERATED THE HYDROGEN OKAY SO THAT'S AN ACID PLUS A METAL NOW AN ACID PLUS A BASE WHICH IS THIS ONE WE'VE JUST DONE A METAL OXIDE THE METAL OXIDES ARE BASES YOU CAN THINK OF THEM AS BEING ALKALINE WE CALL IT BASIC BUT VERY VERY SIMILAR SORT OF THING TO ALKALINE OKAY SO WHAT HAPPENS WITH A BASE AND AN ACID] WITH+ A- BASE+ AND AN+ ACID+
22.242 [HOW ABOUT THIS ONE ONE YOU TRY THIS ONE ON YOUR OWN ONE THIRD TAKE AWAY ONE TWELFTH] ONE+ THIRD+ ONE+ THIRD+ TAKE+ AWAY+
22.151 [BURIED IN SOME] BURIED+ BURIED+
21.824 [SHIFT F SEVEN] SHIFT+ F+ SEVEN+
21.823 [NO SUGAR NO MILK] NO+ SUGAR+ NO+ MILK+
21.639 [IT'LL BE DR] DR+ DR+
21.611 [BAR NINE IS BAR ONE AN OCTAVE LOWER HERE'S THE BIT THAT'S IMPORTANT THIS IS BAR NINE AN OCTAVE LOWER OKAY NOW THEN WILL YOU PLEASE COPY PRECISELY WHAT IS THERE AT BAR TEN THE MUSIC YOU NEED ONE BAR OF MUSIC LINE WITH THOSE BLOBS WHICH ARE THE NOTE HEADS IN EXACTLY THE RIGHT PLACES JUST COPY WHAT'S IN THE BOOK] COPY+ WHAT+ BAR+ TEN+
21.598 [CHOKE KIND OF CHOKE DOWN] KIND+ OF- CHOKE+ DOWN+
21.457 [POWER OF THE UNCONSCIOUS] POWER+ OF+ THE UNCONSCIOUS+

21.397 [GARDEN TOOLS] GARDEN+ TOOLS+
21.394 [OKAY EIGHT TIMES ONE EIGHT LOTS OF ONE EIGHT ONES] EIGHT+ ONES+ EIGHT+
21.366 [PROJECT ENGINEER] PROJECT+ ENGINEER+
21.221 [IT COMES TO SIX WHAT DOES IT MEAN WHY WOULD WE WANT TO WORK OUT THREE TIMES TWO WELL LET'S SAY IF YOU HAD TWO PS IF YOU HAD THREE TWO PS YOU MIGHT WANT TO WORK OUT HOW MUCH THAT COMES TO HOW MUCH WOULD IT COME TO] HOW+ MUCH- WOULD+ THREE+ TWO+ PS+ COME+ TO+
21.179 [TERMS OF ENGAGEMENT] TERMS+ OF- ENGAGEMENT+
21.171 [LETTER] LETTER+ LETTER+
20.871 [AH WELL I FORGOT THE NAME OF IT THE ST NICHOLAS] NICHOLAS+ NICHOLAS+
20.825 [PLAN PRINTING WE'RE AGAIN] PLAN+ PRINTING+
20.490 [YOU MEAN FROM OBJECTORS] FROM+ OBJECTORS+
20.461 [BUY A BOTTLE OF WHISKY AND ORDER UP A HAGGIS] WHISKY+ AND- HAGGIS+

From Test Set 1

44.576 [THIS ONE IS CALLED NONNIE] NONNIE+ NONNIE+ NONNIE+
37.415 [ARE THESE NATIONAL ACCOUNT OR KEY ACCOUNT] KEY+ ACCOUNT+ KEY+ ACCOUNT+
33.060 [REAL GEM] REAL+ GEM+ GEM+
32.573 [HIGH TENSILE STEEL] HIGH+ TENSILE+ STEEL+
31.881 [NONNIE NONNIE NONNIE] NONNIE+ NONNIE+
31.033 [INTER PERSONAL SKILLS] INTER+ PERSONAL+ SKILLS+
30.686 [AKSED SAY AGAIN] AKSED+ AKSED+
30.656 [GETTING YOUR POINT ACROSS EFFECTIVELY YEAH] GETTING+ YOUR+ POINT+ ACROSS+ EFFECTIVELY+
30.456 [YEAH YEAH BUT WHY THE GRAPHICS IS JUST SO YOU CAN DO SCREEN DUMPS ISN'T IT] SCREEN+ DUMPS+ SCREEN+
29.578 [AYE THAT IT'D BE THE VOLTAROL THAT WOULD GIVE HER THE THE BLACK STUFF COMING THROUGH] BLACK+ STUFF+ AYE+ COMING+ THROUGH+ AYE+ AYE+
29.505 [OF PRIDE BEFORE A FALL] PRIDE+ BEFORE+ A+ FALL+
28.186 [IT WAS WAS IT AN OLD ENGLISH SHEEPDOG] OLD+ ENGLISH+ SHEEPDOG+
27.783 [TWENTY MILLION COLLEGE GOES UP IN FLAMES BACON'S BURNING COME EVALUATION WHICH DO YOU THINK THE MOST EFFECTIVE SO FAR REPEAT WHAT WAS IT GOODNESS GRACIOUS GREAT BALLS OF FIRE FIRE FIRE BACON'S COLLEGE ABLAZE IT'S HOT IN THE KITCHEN AND YOURS WERE] TWENTY+ MILLION+ COLLEGE+ GOES+ UP- IN+ FLAMES+
27.177 [IT'S PART OF THE THIRTY SIX HECTARES CHAIRMAN BUT IT'S ONLY A SEVEN HECTARE SITE] SEVEN+ HECTARE+ SITE+

27.109 [AND ALSO OF COURSE THE FACT THAT THE GERMAN MARKET HAD CLOSED] THE+ GERMAN+ MARKET+ HAD+ CLOSED+
27.105 [YEAH OKAY AFTER THE FIRST FIVE NUMBERS THAT'S IT BECAUSE WHAT YOU WELL YOU EITHER HOLD ON TO THE FIRST FIVE OR SIX AND THEN YOU LOSE THE REST OR SOMETIMES YOU REMEMBER THE BEGINNING AND THE END AND YOU LOSE THE BIT IN THE MIDDLE AH IT'S LIKE THAT GAME THAT THEY USED TO PLAY ON CRACKERJACK FOR THOSE OF YOU OLD ENOUGH TO REMEMBER CRACKERJACK] CRACKERJACK+ CRACKERJACK+
26.639 [WAIVER OF PREMIUM] WAIVER+ OF+ PREMIUM+
26.303 [BUT YOU DON'T HAVE A COSTING YOU DO HAVE A COSTING QUOTE FORM] COSTING+ QUOTE+ FORM+
26.090 [NON PECUNIARY] NON+ PECUNIARY+
25.160 [IT'S TODAY'S GUARDIAN] IT'S+ TODAY'S+ GUARDIAN+
25.105 [I SWEAR BY ALMIGHTY GOD] I+ SWEAR+ BY+ ALMIGHTY+ GOD+
24.970 [YEAH THEY'RE SHOOTING AT US IN MOUNT CARMEL] MOUNT+ CARMEL+
24.967 [GOLF ROMEO ALPHA ALPHA ROMEO] ALPHA+ ROMEO+
24.832 [YES ADAPTING FOR CHANGE] ADAPTING+ FOR+ CHANGE+
23.970 [THEY THEY DROVE GENERATORS] THEY+ DROVE+ GENERATORS+
23.801 [IT DIDN'T SAY TO PUT SMART DRIVE AFTER ANSI DID IT IT SAID TO PUT ANSI BEFORE WINDOWS OR KEYBOARD BEFORE WINDOWS] KEYBOARD+ BEFORE+ WINDOWS+
23.445 [MR CURTIS] MR+ CURTIS+
22.929 [I'VE A FEELING B E T N.] B+ E+ T+ N.+
22.480 [ON THE Y AXIS] ON+ THE- Y+ AXIS+
22.373 [THAT WAS YOUR CANDY PEEL] CANDY+ PEEL+
22.252 [MR SPITTLE] MR+ SPITTLE+
21.923 [TWENTY METRES PER SECOND PER SECOND] TWENTY+ METRES+ PER+ SECOND- PER+ SECOND+
21.745 [MY LORD I CALL SERGEANT TAKE THE BOOK IN YOUR HAND AND REPEAT AFTER I SWEAR BY ALMIGHTY GOD] I+ SWEAR+ BY+ ALMIGHTY+ GOD-
21.735 [I'M A TECHNICAL AUTHOR] TECHNICAL+ AUTHOR+
21.600 [NOT IN STENNESS] NOT+ IN+ STENNESS+
21.590 [LORD I CALL SUPERINTENDENT PLEASE TAKE THE BOOK IN YOUR HAND I SWEAR BY ALMIGHTY GOD] I+ SWEAR+ BY+ ALMIGHTY+ GOD-
21.369 [ANCIENT GREEK] ANCIENT+ GREEK+
21.363 [THERE'S THE TWO RONS] THE+ TWO+ RONS+
21.318 [MM I HAVEN'T GOT THAT AND OF COURSE I COULDN'T RUN YOUR VERSION OF SETVER UNLESS I HAD SETVER COS IT WOULD SAY INCORRECT DOS VERSION] INCORRECT+ SETVER+
21.298 [PER AD] PER+ AD+
21.162 [SAY ASKED] ASKED+ ASKED+
21.071 [OH NO IN THE IN THE FIFTIES WE WERE ON THE THE GLASS BOWL FITTINGS YES] ON+ THE- GLASS+ BOWL+ FITTINGS+

20.995 [I'M ANNOYED ABOUT THAT YOU KNOW WE NEGOTIATED ONE PRICE DIDN'T GET AND THEN WE WE'RE SENDING THEM THE NET LIST AND SAYING THERE'S YOUR NEW PRICE] SENDING+ THEM- THE+ NET+ LIST+

20.881 [IT'S PART A TRIAL BUNDLE] TRIAL+ BUNDLE+

20.863 [MY LORD NEXT RAISE THE BOOK IN YOUR RIGHT HAND I SWEAR BY ALMIGHTY GOD] I- SWEAR+ BY+ ALMIGHTY+ GOD-

20.717 [GREEN OAR] GREEN+ OAR+

20.620 [AMUSE KIDS] AMUSE+ KIDS+

20.550 [DID HAVE TABLE LAMPS] TABLE+ LAMPS+

20.511 [THE FACT OF THE MATTER IS THAT AT THE NEXT GENERAL ELECTION THE STORMANT OR THE WESTMINSTER ELECTION GERRY ADDAMS WILL WIN THAT SEAT NOT BECAUSE HE CARRIED THAT COFFIN BUT BECAUSE THE PEOPLE ON HIS SIDE ARE RAVAGED THEIR NERVE ENDS ARE TORN RAW AND IF HE BACKED OFF AND DIDN'T SUPPORT A DEAD FELLOW IRISHMAN HE'D BE DEAD POLITICALLY] IS+ GERRY+ ADDAMS+

20.252 [I THINK THERE'S ONLY THERE'S ONLY THE BED THERE LEFT NOW] THERE'S+ ONLY+ THE+ BED+ THERE+ LEFT+

20.198 [YES ALL QUITE TIDY] ALL+ QUITE+ TIDY+

20.197 [EVERYTHING WAS DONE BY HAND] EVERYTHING+ WAS+ DONE+ BY+ HAND+ EVERYTHING+

20.151 [THAT THE EVIDENCE I SHALL GIVE] THAT+ THE+ EVIDENCE+ I+ SHALL+ GIVE+

20.151 [THAT THE EVIDENCE I SHALL GIVE] THAT+ THE+ EVIDENCE+ I+ SHALL+ GIVE+

20.151 [THAT THE EVIDENCE I SHALL GIVE] THAT+ THE+ EVIDENCE+ I+ SHALL+ GIVE+

20.151 [THAT THE EVIDENCE I SHALL GIVE] THAT+ THE+ EVIDENCE+ I+ SHALL+ GIVE+

20.151 [THAT THE EVIDENCE I SHALL GIVE] THAT+ THE+ EVIDENCE+ I+ SHALL+ GIVE+

20.151 [THAT THE EVIDENCE I SHALL GIVE] THAT+ THE+ EVIDENCE+ I+ SHALL+ GIVE+

20.126 [SEARCH LIGHTS] SEARCH+ LIGHTS+

19.910 [MARVELLOUS TRACEY ABSOLUTELY MARVELLOUS PUT THE NEXT BIT ON FOR ME SAVE ME THE JOB SAY NOTTINGHAMSHIRE'S BIG BANG] NOTTINGHAMSHIRE'S+ BIG+ BANG+

19.697 [PER POLICY] PER+ POLICY+

19.683 [LIKELY TO BE REJECTED] LIKELY+ TO- BE- REJECTED+

19.532 [NO PROBLEM STARTING TEMPERATURE IS PLUS EIGHT ONE ZERO ONE NINE] ONE+ ZERO+ ONE+ NINE+

19.426 [RIGHT I'M ROSEMARY I AM A TECHNICAL LEADER AT MANAGEMENT SERVICES IN WHICH IS A A FAIRLY NEW ROLE] A+ TECHNICAL+ LEADER+

19.401 [I I MISREAD PREVIOUS THE WORD PREVIOUS] WORD+ PREVIOUS+

19.385 [JUST IS IT LESMAHAGOW IS IT] IT+ IS+ LESMAHAGOW+

19.263 [SHALL WE CALL ALAN TURNER] ALAN+ TURNER+
19.006 [MHM AND THIS WAS JUST AN INNOCENT PARTY] PARTY+ MHM+ MHM+ MHM+ MHM+ MHM+
19.004 [OH WRITERS NEWS YEAH THE OTHER ONE] WRITERS+ NEWS+
18.853 [IT LASTS IT LASTS EIGHT MINUTES CHAIRMAN] IT+ LASTS+ EIGHT+ MINUTES+
18.833 [NO I'M A FIRE OFFICER] I'M+ A+ FIRE+ OFFICER+
18.827 [NOT IN THE COOPERATIVES] NOT+ IN+ THE- COOPERATIVES+
18.731 [IT'S A MOORHEN] A+ MOORHEN+
18.731 [DO A LITERAL] A+ LITERAL+
18.534 [AND IT'S MADE OF FIBRE GLASS IS IT] FIBRE+ GLASS+
18.533 [SECONDHAND NO NO NO NEW FURNITURE] NO+ NEW+ FURNITURE+
18.209 [WORKING IN THE QUARRIES LOADING THE LORRIES WITH A HAND SHOVEL] HAND+ SHOVEL+
18.204 [IF YOU'RE STUCK FOR SOMEWHERE TO TAKE THE KIDS ON FIREWORK NIGHT AND YOU WANT TO KNOW THE ONE THAT'S NEAREST TO YOU ACTION LINE HAVE GOT A GREAT BIG LONG LIST YOU'VE GOT PILES HAVEN'T YOU] PILES+ AND- PILES+

Appendix B

"Learning on the Job" :

The Application of Machine Learning within the Speech Recognition Decoder

Paper presented at 2001 Workshop on Innovation in Speech Processing - WISP 2001
and published in Proceedings of the Institute of Acoustics, Vol. 23 (3), pp 71-79.

Institute of Acoustics Workshop on Innovation in Speech Processing, Stratford-on-Avon, May, 2001

LEARNING ON THE JOB: THE APPLICATION OF MACHINE LEARNING WITHIN THE SPEECH DECODER

M A Huckvale Department of Phonetics and Linguistics, University College London, Gower Street, London WC1E 6BT, U.K.
G J A Hunter Department of Phonetics and Linguistics, University College London, Gower Street, London WC1E 6BT, U.K.

1. INTRODUCTION

The current approach to the training of large vocabulary continuous speech recognition (LVCSR) systems involves the use of large corpora of text and labelled audio recordings [1]. These resources are analysed and statistics extracted so that the recogniser can determine the likelihood that the observed signal would have been generated from each sentence it proposes. The analysis of corpora is performed *off-line*, using statistical language models of the text (often trigram models of words), and acoustic models of the signal (often hidden Markov models of phonemes).

To this off-line processing, recent years have seen a growth in the use of methods of adaptation in which the general statistical models are tuned to the specific characteristics of a given speaker, a given acoustic environment or a given topic. These adaptation processes modify the stored characteristics of the language model and the acoustic model to improve the probability that the correct interpretation would have given rise to the observed signal.

No one would argue that these components provide a perfect model of the true statistical distribution of words and sounds. Weaknesses in typical acoustic models include:

- crude modelling of the interdependencies between the acoustic forms of different phones
- no model of systematic pronunciation variation across different contexts or speakers
- little exploitation of durational or pitch cues
- no exploitation of knowledge of style, emotion, or physiological state of the speaker

Weaknesses in typical language models include

- restriction to short-distance dependencies within sentence (trigram models)
- little exploitation of topic or meaning or grammaticality
- poor predictive power for novel or rare events
- limited vocabulary and inability to deal with novel words

These weaknesses are, of course, opportunities for research; and much effort has been spent at looking at these.

There are also many weaknesses that can be seen within the decoder: how these statistical components are exploited in recognition. Weaknesses here include:

- arbitrary balancing of probabilities between the acoustic and language models
- ignorance of interactions between the acoustic model and the language model
- assumptions that words don't overlap in time
- inability to deal with disfluencies and restarts

These are less common areas for research.

Thus we arrive at the present situation in which work is required on many fronts, but each aspect may in itself only provide a modest improvement in performance. It is as if there are many small weaknesses rather than one significant problem. A serious consequence of this situation in speech recognition research is that workers on one small aspect do not know what effect their 'improvements' will have in combination with the work of others. We have been working in the area of morphology for speech recognition [2] but we do not know whether the improvements we've seen will show up in combination with more sophisticated language models or with state-of-the-art acoustic models.

In this paper we are looking towards a 'third way'. Rather than try to build better statistical models, or try to find ways of adapting them to the context, we seek to apply general machine learning principles within the decoder. Thus the decoder will monitor and modify its own behaviour by 'learning on the job'. This work is very much in the exploratory stage. We do not yet know whether the approach will make any significant impact. We do not yet know how it relates to other work in improving language models and acoustic models. We do not even know the best way to make it work.

Our learning decoder is able to relate the correct transcription of an utterance to the complete list of hypotheses that it generated during its attempt at decoding the signal. By looking at the correct and incorrect hypotheses over large numbers of training utterances, it tries to find features of these hypotheses that correlate with their correctness (or with their incorrectness). The aim is not to replace the language model or acoustic model, nor to act as an alternative to adaptation. Instead the machine learning should identify and compensate for common errors made during decoding. Those features that correlate with *correct* can be used to improve the score of probably correct hypotheses, and those features that correlate with *incorrect* can be used to worsen the score of probably incorrect hypotheses. We can use data held-out from training to evaluate the effect of the learning component.

In this paper, we describe how we have implemented and tested this application of machine learning within the decoder of a large vocabulary continuous speech recognition system. In section 2 we describe the mathematical framework we have adopted, while in section 3 we describe a small experiment proposed only as a proof-of-

concept. In section 4 we reflect on the promises offered by the technique and make suggestions for further investigations.

2. Supervised machine learning in the decoder

The aim of the machine learning system is to

- uncover characteristic features of sentence fragment hypotheses which correlate with the correctness of the hypothesis, and
- deliver a probability to the decoder that a sentence fragment is correct given the features that it exhibits.

We describe each of these in turn.

2.1 Selection of features

What features of a sentence fragment hypothesis would assist in determining its probability of being correct? Any features we choose should be complementary to the information provided by the acoustic model and the language model.

In terms of acoustic information, these features might be based on:

- articulation rate, tempo variations, segment durations
- fundamental frequency, voice quality
- articulatory quality
- level of background noise
- detection of speaker, accent, style, emotion or physiological state

In terms of linguistic information, these features might be based on:

- collocational information about words across whole sentences
- measures of grammaticality
- measures of semantic relationships between words

Although many of these aspects of language are likely to influence how a listener decodes an utterance, it is just very complicated to see how they can all be modelled independently and all incorporated in the decoding.

Worse, in many cases we don't know the relative importance of the different features, not how they interact. It is very hard to judge the *utility* of the information provided by a feature. We may run into the problem highlighted by Rosenfeld [3] that we will never have enough data to model rare events - because they are rare.

Thus the first task of our machine learning component will be to decide which of the very many possible features will be of use in practice. Since it is relatively easy to suggest features, but hard to know how useful they are, we leave this task up to the learning system. We simply suggest a very large number of *possible* features and let the system decide which ones to take note of. A useful measure of utility is *mutual information* [10]. For some binary feature f_i and some correctness indicator y, we can calculate the mutual information between f_i and y as:

$$MI(f_i,y)=\sum_{f=0,1}\sum_{g=0,1} p(f_i=f,y=g)\log(\frac{p(y=g\mid f_i=f)}{p(y=g)}) \quad (1)$$

We can choose features with high mutual information shown between the feature and the known correctness of a hypothesis. Features with high mutual information may be useful in predicting correctness or incorrectness and are saved for evaluation in combination.

2.2 Probability modelling of features

Given some signal S and some hypothesis W, we normally calculate the probability that a hypothesis is an interpretation of a signal using Bayes' theorem

$$p(W|S) = p(S|W).p(W)/p(S) \quad (2)$$

Where p(S|W) is the probability that the hypothesis *generated* the signal calculated by the acoustic model, and p(W) is the probability of the hypothesis itself, as calculated by the language model. The decoder seeks to find the single hypothesis that maximises p(W|S).

To incorporate knowledge about some additional features of a hypothesis F(W) not covered by the language model, we can extend the language model to incorporate the prediction of some property y indicating the correctness of the hypothesis:

$$p'(W,y) = p(W).p(y|F(W)) \quad (3)$$

assuming that the language model and the predictions from the features are independent. The probability that a hypothesis is correct given the features of the hypothesis can be expressed in terms of an *exponential model* of the form

$$p(y=correct\mid F(W))=\frac{\exp[\sum_i \lambda_i f_i]}{1+\exp[\sum_i \lambda_i f_i]} \quad (4)$$

where f_i is 1 if the feature *i* is present in the list F(W). The $\{\lambda_i\}$ are constants found from training data. A particular benefit of this model is that the $\{\lambda_i\}$ can be estimated using the principle of *maximum entropy*. Here the least constraining assumptions are drawn from the training data. The $\{\lambda_i\}$ are found by maximising the entropy function

$$\Psi(\lambda)=-\sum_x p(x)\log(1+\exp[\sum_i \lambda_i f_i])+\sum_i \lambda_i p(f_i) \quad (5)$$

where x refers to each different training pattern, p(x) is the probability that the pattern occurs in the training data, and $p(f_i)$ is the probability that feature *i* is seen. We choose to find the maximum of this function using a method of functional optimisation [4]. Other approaches can be found in [5].

3. Experiment

3.1 Materials

Text material for training and testing was selected from the British National Corpus [6]. 80M word of text was reserved for training, and the rest for testing. The corpus was pre-processed to remove all punctuation except for sentence markers, and to convert all numeric items and abbreviations to whole words. A vocabulary of 65,000 words was generated from the most common words in the training portion.

For this experiment we used 1000 spoken sentences taken from the testing portion of the BNC, 100 each from 5 male and 5 female speakers of British English. These were converted to word lattices using the Abbot system [7] with a 65,000-word pronunciation dictionary adapted from BEEP [8] and supplemented with pronunciations from a letter-to-sound system. Abbot was run with parameters provided by Steve Renals to increase the maximum number of hypotheses considered per node to 100.

A language model was constructed for the 65,000-word lexicon using the 80Mword training portion of the BNC. This was performed using the CMU-Cambridge toolkit [9] using Good-Turing discounting.

Decoding of the word lattices using the language model was performed by the UCL decoder, which is able to report node-by-node the currently considered sentence fragment hypotheses for each time step in the word lattice. These hypotheses always extend from the start of the sentence to a word that ends at the current node. They are marked with an overall log probability found during decoding from the acoustic model and the language model.

3.2 Preparation

The hypotheses produced during the decoding of the 1000 sentences were marked for correctness using the known transcription. For training and testing the maximum entropy feature model, we used only those hypotheses that originated from nodes where a correct answer was present within the top 100 hypotheses. This gave us a total of 430,000 hypotheses, of which 26,000 were correct. On average each hypothesis contained 5.65 words.

10% of the data (10 sentences) was reserved from each speaker for testing; the rest was input to the training procedure.

3.3 Feature generation

For this experiment we based our features simply on the collocational properties of word classes within the hypotheses. To do this we designed a set of 50 word classes using word frequency information generated from the training corpus. The word classes were chosen to have approximately similar frequencies in the training corpus. This was achieved by studying the relative frequency of the 50 most common words and the frequency of the 50 most common BNC word tags. We found that a combination of the 25 most common words, 24 most common tags and 1 miscellaneous class gave a

suitable mapping from each word to one of 50 classes. The list of classes is shown in table 1

Table 1 - Word Classes

Class	Word	Class	Tag	Description
1	THE	26	NN1	Singular Noun
2	<S>	27	MISC	Miscellaneous
3	OF	28	AJ0	General Adjective
4	AND	29	NN2	Plural Noun
5	TO	30	AV0	General Adverb
6	A	31	NP0	Proper Noun
7	IN	32	CRD	Cardinal Number
8	IS	33	PNP	Personal Pronoun
9	THAT	34	DT0	General Determiner
10	WAS	35	VVI	Verb Infinitive
11	FOR	36	PRP	Preposition
12	IT	37	VVN	Past Participle Verb
13	ON	38	VM0	Modal Aux. Verb
14	WITH	39	VVD	Past Tense of Verb
15	AS	40	VVG	Verb (-ing form)
16	HE	41	DPS	Possessive Determiner
17	BE	42	NN0	Noun (not number specific)
18	BY	43	CJS	Subordinating Conjunction
19	AT	44	DTQ	wh- determiner
20	I	45	VVZ	Present form (-s) of verb
21	ONE	46	AT0	"Article" determiner (a, the,an)
22	HIS	47	AJ0-NN1	Word can be noun or adjective
23	NOT	48	VBB	Present tense of verb "to be"
24	BUT	49	AVP	Adverb particle (up, off, ...)
25	FROM	50	VHD	Past tense of verb "to have"

Using these word classes, collocational features were proposed as follows: feature F(m,n) is 1 if and only if word-class m occurs in the hypothesis before word class n. Thus each hypothesis is converted to a (sparse) vector of 2500 bits.

3.4 Feature Winnowing

To determine which of the 2500 features had some potential for predicting the correctness of the hypothesis, a first 'winnowing' stage was implemented using a mutual information criterion as described in section 2.1.

The winnowing procedure looked only at those hypotheses that were either correct or which had a score better than the correct hypothesis on the node. The mutual information was calculated between each feature f_i and the correctness indicator y. The 50 features showing the greatest values were retained for input to the maximum entropy modelling.

3.5 Maximum entropy modelling

From the list of 50 features showing the greatest mutual information, maximum entropy models are made using a greedy algorithm (following [5]) that considers first the best model with one feature, then the best second feature that can be added to the first, the best third feature that can be added to the first two, and so on.

The maximum entropy modelling halts when the additional benefit of adding another feature falls below some threshold. A typical example of a model of 10 features is shown in table 2.

Table 2 - Example Maximum Entropy Model

No.	Feature	Lambda	Description
1	1<26	-1.55945	"the" before singular noun
2	26<26	-1.50821	singular noun before singular noun
3	27<1	-1.42796	miscellaneous before "the"
4	1<28	-1.558	"the" before general adjective
5	30<1	-1.42289	general adverb before "the"
6	30<31	-3.89828	general adverb before proper noun
7	34<1	-1.3658	general determiner before "the"
8	27<8	-1.92252	miscellaneous before "is"
9	1<27	-1.43065	"the" before miscellaneous
10	26<35	-1.56516	singular noun before infinitive

Note that all the lambda values are negative, indicating that these features reduce the likelihood of any hypothesis containing these features being correct. Features that increased the likelihood of a hypothesis being correct were found by the winnowing procedure but they did not find their way into any maximum entropy model.

At first sight these features of incorrect hypotheses do not look particularly odd. However a feature is useful if its frequency of occurrence is different in correct and incorrect hypotheses. Thus the fixation on the use of 'the' may simply indicate that the recogniser is hypothesising this word too often.

3.6 Evaluation

To evaluate the feature selection and maximum entropy models, the 10% of data reserved for testing was processed through the word-class mapping and feature extraction stages. The overall score for each hypothesis was then adjusted using equations (3) and (4) for each of the selected features and calculated lambda parameters found from the 90% of data used for training. The procedure was then repeated 10 times for each possible division between test and training.

To evaluate the effectiveness of the new scores for each hypothesis, we calculated the average rank of the correct answer in the list of hypotheses generated for each node. After rescoring, the hypothesis list was resorted and the average rank of the correct answer recalculated. The results are shown in table 3:

Table 3 - Change in Ranking of First Correct Hypothesis

Test Data Set	Mean Correct Ranking (Before)	Mean Correct Ranking (After)	Mean Improvement
hyp0.lst	13.15	9.94	3.21
hyp1.lst	15.63	12.38	3.25
hyp2.lst	14.96	11.99	2.97
hyp3.lst	16.19	11.89	4.30
hyp4.lst	13.96	10.59	3.37
hyp5.lst	15.19	10.12	5.07
hyp6.lst	17.04	11.06	5.98
hyp7.lst	14.51	10.56	3.95
hyp8.lst	14.21	10.40	3.81
hyp9.lst	14.26	10.16	4.10

Overall the mean ranking of the correct answer improved by 4 places, from an average rank of 14.9 to an average rank of 10.9. The results seem consistent across each rotation of data. We have not yet determined how these improvements in ranking affect word recognition score. For this experiment we simply wanted to show that the maximum entropy model made consistent changes to scores in the right direction.

4. Discussion

The experiment described above is only a first attempt at applying the idea of machine learning within the decoder, and serves only as a proof of concept that the idea holds some promise. We made many arbitrary decisions in feature analysis and in modelling and these can almost certainly be improved.

Now that we have the basic framework for experimentation we would like to look at:

1. choosing word classes on the basis of either grammatical functionality, or on the basis of how the word contributes to meaning
2. choosing other features based on the position of the word with respect to words that become before and after it
3. finding the best way to exploit the modified scores in the decoder: whether the modifications should be actually incorporated with scores from the acoustic model and language model, or whether they should be used simply to help rank hypotheses within a node.
4. determining the effect of the machine learning on word accuracy
5. determining the effect of the machine learning on sentences drawn from a different corpus spoken by speakers outside the training set.

One particular problem that might arise with this technique is that the features found in one set of data fail to be useful in another. On the other hand, the technique trawls through a large number of features to find ones that occur commonly and have the greatest effect. We are hopeful that the technique can be extended and refined to incorporate acoustic as well as linguistic features, and that a general learning framework can be established within the decoder to identify further features automatically.

Acknowledgements

Gordon Hunter is supported by a research studentship from the U.K Engineering and Physical Sciences Research Council. Some of the work reported here was conducted under the EPSRC project "Enhanced Language Modelling", Grant No GR/L81406. Thanks to Alex Fang for help with the BNC corpus and the generation of the language models.

References

[1] S. Young. Large Vocabulary Continuous Speech Recognition: A Review. In *IEEE Signal Processing Magazine*, 13(5), 1996, 45-57.

[2] A.C. Fang, M.A. Huckvale. Enhanced Language Modelling with Phonologically Constrained Morphological Analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5-9 June 2000, Istanbul, Turkey.

[3] R. Rosenfeld. Incorporating Linguistic Structure into Statistical Language Models. In *Philosophical Transactions of the Royal Society of London A*, 358, 2000, 1311-1324.

[4] J.A. Nelder, R. Mead. A simplex method for function minimization. In *The Computer Journal*, vol.7, 1965, The British Computer Society, 308-313.

[5] A. Berger, S. Della Pietra, V. Della Pietra. A maximum entropy approach to natural language processing. In *Computational Linguistics*, 22, 1996, pp1-36.

[6] L. Burnard. *Users Reference Guide for the British National Corpus.* Oxford University Computing Services, Oxford, 1995.

[7] M. Hochberg, S. Renals, A. Robinson. ABBOT: The CUED Hybrid Connectionist-HMM Large Vocabulary Recognition System. In *Proc. Language Technology Workshop,* Austin Texas, Jan 1995. Morgan Kaufmann.

[8] A. Robinson, BEEP Pronunciation Dictionary. Retrieved from World Wide Web: ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz

[9] P. Clarkson, R. Rosenfeld. Statistical Language Modeling using the CMU-Cambridge Toolkit. In *Proc. Eurospeech 97*, 1997.

[10] R. Rosenfeld. A Maximum Entropy Approach to Adaptive Statistical Language Modelling. In *Computer, Speech and Language*, 10, 1996, 187-288.